SFB 649 Discussion Paper 2016-040

Principal Component Analysis in an Asymmetric Norm

Ngoc M. Tran*
Petra Burdejová*²
Maria Osipenko*²
Wolfgang K. Härdle*²

* University of Texas at Austin, United States of America
*² Humboldt-Universität zu Berlin, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

Principal Component Analysis in an Asymmetric Norm*

Ngoc M. Tran[1,2], Petra Burdejová[3], Maria Osipenko[3] and Wolfgang K. Härdle[3,4]

[1]Department of Mathematics, University of Texas at Austin, USA.
[1]Institute for Applied Mathematics, University of Bonn, Germany.
[3]Humboldt-Universität zu Berlin, C.A.S.E. - Center for Applied Statistics and Economics, Unter den Linden 6, Berlin, Germany.
[4]Sim Kee Boon Institute for Financial Economics, Singapore Management University, 90 Stamford Road, 6th Level, School of Economics, Singapore 178903.

Abstract

Principal component analysis (PCA) is a widely used dimension reduction tool in the analysis of high-dimensional data. However, in many applications such as risk quantification in finance or climatology, one is interested in capturing the tail variations rather than variation around the mean. In this paper, we develop Principal Expectile Analysis (PEC), which generalizes PCA for expectiles. It can be seen as a dimension reduction tool for extreme value theory, where one approximates fluctuations in the τ-expectile level of the data by a low dimensional subspace. We provide algorithms based on iterative least squares, prove upper bounds on their convergence times, and compare their performances in a simulation study. We apply the algorithms to a Chinese weather dataset and fMRI data from an investment decision study.

Keywords: principal components; asymmetric norm; dimension reduction; quantile; expectile; fMRI; risk attitude; brain imaging; temperature; functional data
JEL Classification: C38, C55, C61, C63, D81

1 Introduction

Principal component analysis (PCA) and its functional version (FPCA) are widely used for dimension reduction. This method has been successfully applied in many fields such as gene expression measurements, weather, natural hazard, and environment studies, demographics, etc. The monographs of Jolliffe (2004) and Ramsay and Silverman (2005) contain many examples. The basic principle is to find a basis for a k-dimensional affine linear subspace that best approximates the data. If the data points are finite-dimensional vectors, the basis vectors are called principal components, or factors. If the data points are in an infinite-dimensional Hilbert space, the basis functions are called functional principal components. One then views each observation as residual plus a point in this subspace, which is expressed as a vector in $\mathbb{R}^k$ of coefficients, also

*This research was supported by Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk" and the International Research Training Group IRTG 1792 "High Dimensional Non Stationary Time Series". Ngoc Tran was also supported by DARPA (HR0011-12-1-0011) and an award from the Simons Foundation (# 197982 to The University of Texas at Austin).

called loadings. A classic example is the Canadian temperature dataset in Ramsay and Silverman (2005), where they considered temperature curves recorded daily over a year at multiple stations in an area. The premise is that there are only a few factors influencing the temperature across stations, and that the temperature curve from each station is well-approximated on average by a specific linear combinations of these factors.

In classical PCA and FPCA, the optimal k-dimensional subspace is one that minimizes the L_2-norm of the residual. When $k = 0$, this is the mean of the data. Thus, classical (F)PCA decomposes the data around its mean subspace. In fact, much research in the larger field of functional data analysis have focused on the variation around an average pattern, as seen in the monographs Horváth and Kokoszka (2012), Ferraty and Vieu (2006). In many applications such as risk analysis, however, one is not only interested in functional variations around the mean, but rather those around the tail of the data. For example, one may be interested in the extreme phenomena like drought, rainfall, or heat wave. Can one decompose the data around the 99-th quantile, for instance, and produce some 'best' principal component where only 1% of the observations have positive loadings? In the previous temperature data, for example, this principal component can be interpreted as one that influence locations with extreme temperatures.

Note that the above problem is different from finding the 99-th quantile of the loadings in classical PCA. Doing so corresponds to keeping the same PCA-optimal subspace, and translating it so that each component has 1% positive loadings. The principal components are the same; the data's tail is reflected by the loadings. In our setup, one wants to find a low-dimensional subspace that best approximates the data by some tail measure, say, an appropriate analogue of 99-th quantile. In this case, the data's tail is reflected by the principal components. As we shall show in Section 4, only in some special cases do these two methods give the same subspace.

In this paper, we generalize PCA to Principal Expectile Analysis, a method that for a given expectile level τ produces k principal expectile components (PECs) that best decompose the data around its τ-expectile. Classical PCA corresponds to the case $\tau = 0.5$. Expectiles, proposed by Newey and Powell (1987), are natural analogues of quantiles for the mean. While the τ-quantile minimizes asymmetric ℓ_1-error, the τ-level expectile minimizes asymmetric ℓ_2-error.

Expectiles enjoy several advantages over quantiles, including computational efficiency, see Schnabel (2011). It is also more sensitive to extreme values in the data, and thus is preferred in the calculation of risk measures of a financial asset or a portfolio. For instance, value-at-risk (VaR) is commonly used to measure the downside risk, especially in portfolio risk management. Given a predetermined probability level, VaR represents the quantile of the portfolio loss distribution, see Jorion (2000). Since VaR, which is not a coherent measure, merely depends on the probability value and neglects the size of the downside loss, it has been criticized as a risk measure. Alternative risk measures based on expectiles have been investigated, see Kuan et al. (2009) or Daouia et al. (2016).

Our definitions of PECs are related to the principal directions for quantiles of Fraiman and Pateiro-López (2012). These authors are focused on doing classical PCA for quantile level sets. Since the quantile has to be computed in each direction, their definitions can only be explicitly computed in small dimensions in general. In contrast, we focus on using quantiles and expectiles to generalize PCA. Since its conception, Principal Expectiles Analysis have seen numerous applications, mainly in quantifying risks. In climate analysis, Burdejova et al. (in press 2016) looks for trends and critical changing points in the strength of tropical storms in

two different areas over several decades. Analysis considers the wind data observed every 6 hours represented as functional data for several τ-expectile levels. A proposed test based on principal components shows that there is a significant trend in the shape of the annual pattern of upper wind speed levels of hurricanes. In this setup, PECs yield time varying information of storm strength which lies between 'typical' and 'extreme' behavior. This approach can be applied to any environmental data as which can be represented as annual curves which evolve from year to year, such as daily temperature or log-precipitation curves at specific locations.

The second example concerns energy markets; their fair pricing procedure is driven by functions of the extremal of the data distribution; see López Cabrera and Schulz (2016). In the later paper, functional principal components of precomputed tail event curves are used for forecasting of electricity load. Essentially, with a help of defining the "τ-variance", PEC approach could simplify this 2-step methodology into one step only. Even though in case of electricity load we get the similar results, generally one should be careful, especially in case of dependent data, where the condition of weak-dependence is not fulfilled.

We present two other applications. First one analyzes the climate data of daily temperature over last 5 decades for 159 Chinese stations. This is an analogue to the commonly known approach of Ramsay and Silverman (2005). However, we show that PECs significantly differ from PCs. Further, we observe that while the first component shows the long-term behaviour, the second component is also crucial and corresponds to temporal seasonal extremes. The second application demonstrates the usefullness of PEC in a specific neurobilogical task. Recently, via the RPID (Risk Perception in Investment Decision) experiment data Majer et al. (2015) found strong relations between fRMI reactions and diagnosed risk perception. Empirical results show that one can predict the risk perception parameter of each individual better based on the principal components of the fMRI data. However, their work analyses only the average brain reactions. In other words, we devise to analyse if extreme fMRI reactions can correspond to more extreme behaviors against risk and show that one can have better results for higher level of τ=0.6 than for a commonly taken $\tau = 0.5$, which corresponds to classical PCA.

Our paper is organized as follows. In Section 2 we review quantiles, expectiles and PCA. We then discuss the issues in generalizing PCA to expectiles, and propose a definition for principal expectile components, PrincipalExpectile algorithm and two other variations named TopDown and BottomUp. In Section 4, we prove statistical properties of these estimators. In Section 5, we provide algorithms to compute PEC, TopDown and BottomUp based on iterative weighted least squares, and prove upper bounds on their convergence times. We compare their performances in a simulation study in Section 6. In Section 7, we show an application to a Chinese weather dataset and fMRI data. The last section summarizes our findings.

2 Background

2.1 Quantiles and Expectiles

Let us assume that the data dimension p is fixed. For $y \in \mathbb{R}^p$, we define $y_+ \stackrel{\text{def}}{=} \max(0, y)$, $y_- \stackrel{\text{def}}{=} \max(0, -y)$ coordinatewise. For $\tau \in (0, 1)$, let $\|\cdot\|_1$ denote the L_1-norm in $\mathbb{R}^p$, that is,

$\|y\|_1 = \sum_{j=1}^p |y_j|$. The L_1-norm with weight τ in $\mathbb{R}^p$ is

$$\|y\|_{\tau,1} = \tau\|y_+\|_1 + (1-\tau)\|y_-\|_1 = \sum_{j=1}^p |y_j| \cdot \{\tau I(y_j \geq 0) + (1-\tau)I(y_j < 0)\},$$

where $I(\cdot)$ is the indicator function. Similarly, let $\|\cdot\|_2$ denote the L_2-norm in $\mathbb{R}^p$, $\|y\|_2^2 = \sum_{j=1}^p y_j^2$. The asymmetric L_2-norm with weight τ in $\mathbb{R}^p$ is

$$\|y\|_{\tau,2}^2 = \tau\|y_+\|_2^2 + (1-\tau)\|y_-\|_2^2.$$

When $\tau = 1/2$, we recover constant multiples of the L_1 and L_2-norms, respectively. These two families of norms belong to the general class of asymmetric norms with sign-sensitive weights. These have appeared in approximation theory, see Cobzaş (2013). Some properties we use in this paper are the fact that these norms are convex, and their unit balls restricted to a given orthant in $\mathbb{R}^p$ are weighted simplices for the $\|\cdot\|_{\tau,1}$ norm, and axis-aligned ellipsoids for the $\|\cdot\|_{\tau,2}$ norm. In other words, they coincide with the unit balls of axis-aligned weighted L_1 and L_2 norms.

Let $Y \in \mathbb{R}^p$ be a random variable with cumulative distribution function (cdf) F. The τ-quantile $q_\tau(Y) \in \mathbb{R}^p$ of F_Y is the solution to the following optimization problem

$$q_\tau(Y) = \operatorname*{argmin}_{q \in \mathbb{R}^p} \mathsf{E}\|Y - q\|_{\tau,1}.$$

Similarly, the τ-expectile $e_\tau(Y) \in \mathbb{R}^p$ of F_Y is the solution to

$$e_\tau(Y) = \operatorname*{argmin}_{e \in \mathbb{R}^p} \mathsf{E}\|Y - e\|_{\tau,2}^2.$$

By Cobzaş (2013), the solution exists and is unique, assuming that $\mathsf{E}(Y)$ is finite. This definition guarantees that the τ-quantile $q_\tau(Y)$ is unique even when the cdf F is not invertible. When F is invertible with inverse function F^{-1}, $q_\tau(Y)$ coincides with $F^{-1}(\tau)$, see Cobzaş (2013).

2.2 Classical principal component analysis

There are multiple, equivalent ways to define classical PCA, which generalize to different definitions of principal components for quantiles and expectiles. We focus on two formulations: minimizing the residual sum of squares, and maximizing the variance captured. For further details, see Jolliffe (2004).

Suppose we observe n vectors $Y_1, \ldots, Y_n \in \mathbb{R}^p$ with empirical distribution function (edf) F_n. Write Y for the $n \times p$ data matrix. PCA solves for the k-dimensional affine subspace that best approximates $Y_1, \ldots, Y_n$ in L_2-norm. In matrix terms, we are looking for the constant $m^* \in \mathbb{R}^p$ and the matrix E_k^*, the rank-k matrix that best approximates $Y - \mathbf{1}(m^*)^\top$ in the Frobenius norm. That is,

$$(m_k^*, E_k^*) = \operatorname*{argmin}_{m \in \mathbb{R}^p, E \in \mathbb{R}^{n \times p}: rank(E) = k} \|Y - \mathbf{1}m^\top - E\|_{1/2,2}^2. \tag{1}$$

As written, m is not well-defined: if (m, E) is a solution, then $(m + c, E - \mathbf{1}c^\top)$ is another solution for any c in the column space of E. Geometrically, this means we can express the affine subspace $m + E$ with respect to any chosen point m. It is intuitive to choose m to be the best constant in this affine subspace that approximates Y. By a least squares argument, the solution is $m_k^* = \mathsf{E}(Y)$. That is, it is independent of k and coincides with the best constant approximation to Y. Thus, it is sufficient to assume $\mathsf{E}(Y) = m \equiv 0$, and consider the optimization problem in (1) without the constant term.

Suppose Y is full rank and the eigenvalues of its covariance matrix are all distinct. This is necessary and sufficient for principal components to be unique. Again by least squares argument, for $1 \leq k < p$, one can show that

$$E_k^* \subset E_{k+1}^*, \tag{2}$$

and $E_{k+1}^* - E_k^*$ is the optimal rank-one approximation of $Y - E_k^*$. This has two implications. Firstly, there exists a natural basis for E_k^*. Indeed, there exists a unique ordered sequence of orthonormal vectors $v_1^*, v_2^*, \ldots, v_p^* \in \mathbb{R}^p$ such that $E_1^* = U_1 V_1^\top$, $E_2^* = U_2 V_2^\top$, and so on, where the columns of V_k are the first k v_i^*'s. The v_i^*'s are called the *principal components*, or *factors*. For fixed k, V_k is the *component*, or *factor matrix*, and U_k is the *loading*. The second implication of (2) is that one can compute the principal components by a greedy algorithm which solves k iterations of the one-dimensional version of (1).

The one-dimensional version of (1) has another characterization. The first principal component v^* is the unit vector in $\mathbb{R}^p$ which maximizes the variance of the data projected onto the subspace spanned by v^*. That is,

$$v^* = \operatorname*{argmax}_{v \in \mathbb{R}^p, v^\top v = 1} \operatorname{Var}\{vv^\top Y_i : 1 \leq i \leq n\} = \operatorname*{argmax}_{v \in \mathbb{R}^p, v^\top v = 1} n^{-1} \sum_{i=1}^{n} (v^\top Y_i - \overline{v^\top Y})^2, \tag{3}$$

where $\operatorname{Var}\{\cdot\}$ is the variance of the sequence in the argument, while $\overline{v^\top Y} = n^{-1} \sum_{i=1}^n v^\top Y_i$ is the mean of the projected data, or equivalently, the projection of the mean $\bar{Y}$ onto the subspace spanned by v. Given that the first principal component is v_1^*, the second principal component v_2^* is the unit vector in $\mathbb{R}^p$ which maximizes the variance of the residual $Y_i - (v_1^*)^\top \bar{Y} - v_1^*(v_1^*)^\top Y_i$, and so on. In this formulation, the data does not have to be pre-centered. The sum $(v_1^*)^\top \bar{Y} + (v_2^*)^\top \bar{Y} + \ldots + (v_k^*)^\top \bar{Y}$ is the overall mean $\bar{Y}$ projected onto the subspace spanned by the first k principal components.

For the benefit of comparisons to Theorem 4.2, let us reformulate (3) as follows. Define

$$C = n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})^\top. \tag{4}$$

Then v^* is the solution to the following optimization problem.

$$\begin{aligned} &\text{maximize } v^\top C v \\ &\text{subject to } v^\top v = 1. \end{aligned}$$

It is clear from this formulation that this optimization problem has a solution unique up to sign if and only if C has a unique largest eigenvalue. For this reason, we shall implicitly assume that all eigenvalues of C are unique.

3 Principal Expectile Analysis

We now generalize the above definitions of PCA to those for expectiles, leading to principal expectile analysis. While we exclusively focus on expectiles in this paper, we note that the generalization for quantiles follows similarly, and algorithms for L_1 matrix factorization can also be adapted to this case.

The two views of PCA, minimizing-least-squares in (1), and maximizing-projected-variance in (4), are no longer equivalent when one optimizes these functions under the asymmetric L_2-norm. This is because the asymmetric norm is not a projection. The analogue of (1) is the following low-rank matrix approximation problem

$$(m_k^*, E_k^*) = \underset{m\in\mathbb{R}^p, E\in\mathbb{R}^{n\times p}: rank(E)=k}{\operatorname{argmin}} \|Y - \mathbf{1}m^\top - E\|_{\tau,2}^2. \tag{5}$$

Again, we may define m to be the best constant approximation to Y on the affine subspace determined by (m, E). For a fixed affine subspace, such a constant is unique, and is the coordinatewise τ-expectile of the residuals $Y - E$. However, the expectile is not additive for $\tau \neq 1/2$. Thus in general, the column space of E_k^* is not a subspace of the column space E_{k+1}^*, the constant m_k^* depends on k, and is not equal to the τ-expectile $e_\tau(Y)$. In other words, even when E_k^* is a well-defined subspace, it does not come with a natural basis, and hence there are no natural candidates for 'principal components'.

To define principal expectile components, one can furnish E_k^* with two types of basis, which we call TopDown and BottomUp. In TopDown, one first finds E_k^*. Then for $j = 1, 2, \ldots, k-1$, one finds E_j, the best j-dimensional subspace approximation to $Y - m_k^*$, subjected to $E_{j-1} \subset E_j \subset E_k^*$. This defines a nested sequence of subspace $E_1 \subset E_2 \subset \ldots \subset E_{k-1} \subset E_k^*$, and hence a basis for E_k^*, such that E_j is an approximation of the best j-dimensional subspace approximation to $Y - m_k^*$ contained in E_k^*. In BottomUp, one first finds E_1^*. Then for $j = 2, \ldots, k$, one finds (m_j, E_j), the optimal j-dimensional affine subspace approximation to Y, subjected to $E_{j-1} \subset E_j$. In each step we re-estimate the constant term. Again, we obtain a nested sequence of subspaces $E_1^* \subset E_2 \subset \ldots \subset E_k$, and constant terms $m_1, \ldots, m_k$, where (m_j, E_j) is the best affine j-dimensional subspace approximation to Y.

When $\tau = 1/2$, that is, when doing usual PCA, both definitions correctly recover the principal components. For $\tau \neq 1/2$, they can produce different output. Interestingly, both in simulations and in practice, their outputs are not significantly different (see Sections 6 and 7). See Section 5 for a formal description of the TopDown and BottomUp algorithms and computational bounds on their convergence times.

Generalization of (3) is more fruitful, both theoretically and computationally. First we need a weighted definition of the variance. Let $Y \in \mathbb{R}$ be a random variable with cdf F. Its *τ-variance* is

$$\mathrm{Var}_\tau(Y) = \mathsf{E}\|Y - e_\tau\|_{\tau,2}^2 = \min_{e\in\mathbb{R}} \mathsf{E}\|Y - e\|_{\tau,2}^2,$$

where $e_\tau = e_\tau(Y)$ is the τ-expectile of Y. When $\tau = 1/2$, this reduces to the usual definition of variance. The direct generalization of (3) is

$$v_\tau^* = \underset{v\in\mathbb{R}^p, v^\top v=1}{\operatorname{argmax}} \operatorname{Var}_\tau\{v^\top Y_i : 1 \le i \le n\} \tag{6}$$

$$= \underset{v\in\mathbb{R}^p, v^\top v=1}{\operatorname{argmax}} n^{-1}\sum_{i=1}^{n}(v^\top Y_i - \mu_\tau)^2 w_i \tag{7}$$

where $\mu_\tau \in \mathbb{R}$ is the τ-expectile of the sequence of n real numbers $v^\top Y_1, \dots v^\top Y_n$, and

$$w_i = \tau \text{ if } \sum_{j=1}^{p} Y_{ij}v_j > \mu_\tau, \text{ and } w_i = 1-\tau \text{ otherwise.} \tag{8}$$

Definition 3.1. Suppose we observe $Y_1, \dots, Y_n \in \mathbb{R}^p$. The first *principal expectile component* (PEC) v_τ^* is the unit vector in $\mathbb{R}^p$ that maximizes the τ-variance of the data projected on the subspace spanned by v_τ^*. That is, v_τ^* solves (7).

Like in classical PCA, the other components are defined based on the residuals, and thus by definition, they are orthogonal to the previously found components. Therefore one obtains a nested sequence of subspace which captures the tail variations of the data.

The τ-variance measures the spread of the data relative to the τ-expectile e_τ. For τ very close to 1, for example, observations above e_τ receives the very high weight τ, while those below receives very little weight. Similarly, for τ very close to 0, observations below e_τ receives most of the weight. In other words, the τ-variance is dominated by the observations more extreme than e_τ. Thus, PEC, the direction that maximizes that τ-variance of the projected data, can be interpreted as the direction with the most 'extreme' behavior in the loadings.

Generalizing principal components to quantiles via its interpretation as variance maximizer is not new. Fraiman and Pateiro-López (2012) define the first principal quantile direction ψ to be the one that maximizes the L_2 norm of the τ-quantile of the centered data, projected in the direction ψ. That is, ψ is the solution of

$$\max_{v\in\mathbb{R}^p : v^\top v=1} \|v^\top q_\tau(Y - \mathsf{E}Y)\|_{1/2,2}.$$

Their definition works for random variables in arbitrary Hilbert spaces. Kong and Mizera (2012) proposed the same definition but without centering Y at $\mathsf{E}Y$. These authors are focused on doing classical PCA for quantile level sets in small dimensions. In contrast, we focus on using expectiles to generalize PCA.

For ease of comparison with Fraiman and Pateiro-López (2012) and the related literature, we give the quantile analogue of our definition of PEC. By replacing the $\|\cdot\|_{\tau,2}^2$ norm with the $\|\cdot\|_{\tau,1}$ norm, one can define the analogue of principal component for quantiles. The analogue of τ-variance is the τ-deviation

$$\operatorname{Dev}_\tau(Y) = \mathsf{E}\|Y - q_\tau(Y)\|_{\tau,1} = \min_{q\in\mathbb{R}} \mathsf{E}\|Y - q\|_{\tau,1}.$$

This leads to the optimization problem

$$v^*_{\tau,L_1} = \underset{v\in\mathbb{R}^p:\sum_j |v_j|=1}{\operatorname{argmax}} \operatorname{Dev}_\tau\{v^\top Y_i : 1 \le i \le n\}.$$

One can define the first *principal quantile component* (PQC) v^*_{τ,L_1} as the L_1-unit vector in $\mathbb{R}^p$ that maximizes the τ-deviation captured by the data projected on the subspace spanned by v^*_{τ,L_1}.

Like the definition of Fraiman and Pateiro-López (2012), one can generalize PEC to the case where Y is a variable in an infinite-dimensional Hilbert space $\mathcal{H}$ by replacing the set $v \in \mathbb{R}^p, v^\top v = 1$ with the unit ball in $\mathcal{H}$. Furthermore, our definition of PEC satisfies many 'nice' properties, some of which are shared by the principal directions of Fraiman and Pateiro-López (2012). For example, the PEC coincides with the classical PC when the distribution of Y is elliptically symmetric, see Proposition 4.2.

4 Statistical properties of PEC

We now show that our definition of PEC satisfies many important properties, such as being compatible to orthogonal transformation of the data, and coinciding with classical PC for elliptically symmetric distributions (cf. Proposition 4.2). More important, we show that the empirical estimator in (7) is consistent under some mild uniqueness assumptions akin to the unique leading eigenvalue assumption in classical PCA.

Proposition 4.1 (Properties of τ-variance)**.** *Let $Y \in \mathbb{R}$ be a random variable. For $\tau \in (0,1)$, the following statements hold:*

- $\operatorname{Var}_\tau(Y + c) = \operatorname{Var}_\tau(Y)$ *for* $c \in \mathbb{R}$,
- $\operatorname{Var}_\tau(sY) = s^2\operatorname{Var}_\tau(Y)$ *for* $s \in \mathbb{R}, s > 0$,
- $\operatorname{Var}_\tau(-Y) = \operatorname{Var}_{1-\tau}(Y)$.

Proof. The first two follow directly from corresponding properties for e_τ. We shall prove that last assertion. Recall that $e_\tau(-Y) = -e_{1-\tau}(Y)$. Thus

$$\begin{aligned}\operatorname{Var}_\tau(-Y) &= \mathsf{E}\|-Y - e_\tau(-Y)\|^2_{\tau,2} = \mathsf{E}\|-\{Y - e_{1-\tau}(Y)\}\|^2_{\tau,2} = \mathsf{E}\|Y - e_{1-\tau}(Y)\|^2_{1-\tau,2} \\ &= \operatorname{Var}_{1-\tau}(Y).\end{aligned}$$

□

As a corollary, we see that PECs are sign-sensitive in general, unless if the distribution of Y is symmetric, or if $\tau = 1/2$.

Corollary 4.1. *For $\tau \in (0,1)$, random variable $Y \in \mathbb{R}^p$, suppose v^*_τ is a first τ-PEC of Y. Then*

$$-v^*_\tau = v^*_{1-\tau},$$

that is, $-v^*_\tau$ is also a first $(1-\tau)$-PEC of Y. Furthermore, if the distribution of Y is symmetric about 0, that is, $Y \overset{\mathcal{L}}{=} -Y$, then $-v^*_\tau$ is also a first τ-PEC of Y.

Proof. By Proposition 4.1, $\operatorname{Var}_\tau(v^{*\top}_\tau Y) = \operatorname{Var}_{1-\tau}\{(-v^{*\top}_\tau)Y\}$. Thus if v^*_τ solves (6) for τ, then $(-v_\tau)^*$ solves (6) for $1-\tau$. If the distribution of Y is symmetric about 0, then

$$\operatorname{Var}_\tau(v^{*\top}_\tau Y) = \operatorname{Var}_{1-\tau}\{v^{*\top}_\tau(-Y)\} = \operatorname{Var}_\tau(v^{*\top}_\tau Y).$$

In this case $-v_\tau^* = v_{1-\tau}^*$ is another τ-PEC of Y. □

Proposition 4.2. *[Properties of principal expectile component] Let $Y \in \mathbb{R}^p$ be a random variable, $v_\tau^*(Y)$ its unique first principal expectile component as given in Definition 3.1.*

1. *For any constant $c \in \mathbb{R}^p$, $v_\tau^*(Y + c) = v_\tau^*(Y)$. In words, the PEC is invariant under translations of the data.*
2. *If $B \in \mathbb{R}^{p\times p}$ is an orthogonal matrix, then $v_\tau^*(BY) = Bv_\tau^*(Y)$. In words, the PEC respects change of basis.*
3. *If the distribution of Y is elliptically symmetric about some point $c \in \mathbb{R}^p$, that is, there exists an invertible $p \times p$ real matrix A such that $BA^{-1}(Y-c) \overset{\mathcal{L}}{=} A^{-1}(Y-c)$ for all orthogonal matrices B, then $v_\tau^*(Y) = v_{1/2}^*(Y)$. In this case, the PEC coincides with the classical PC regardless of τ.*
4. *If the distribution of Y is spherically symmetric about some point $c \in \mathbb{R}^p$, that is, $B(Y - c) \overset{\mathcal{L}}{=} Y - c$ for all orthogonal matrix B, then all directions are principal.*

Proof. By the first part of Proposition 4.1:

$$\begin{aligned}\mathrm{Var}_\tau\{v^\top(Y_i + c) : i = 1, \dots, n\} &= \mathrm{Var}_\tau(v^\top Y_i + v^\top c : i = 1, \dots, n)\\ &= \mathrm{Var}_\tau(v^\top Y_i : i = 1, \dots, n).\end{aligned}$$

This proves the first statement. For the second, note that

$$\mathrm{Var}_\tau(v^\top BY_i : i = 1, \dots, n) = \mathrm{Var}_\tau\{(B^\top v)^\top Y_i : i = 1, \dots, n\}.$$

Thus if v_τ^* is the first τ-PEC of Y, then $(B^\top)^{-1}v_\tau^*$ is the first τ-PEC of BY. But B is orthogonal, that is, $(B^\top)^{-1} = B$. hence Bv_τ^* is the τ-PEC of BY. This proves the second statement. For the third statement, by statement 1, we can assume $c \equiv 0$. Thus $Y = AZ$ where $BZ \overset{\mathcal{L}}{=} Z$ for all orthogonal matrices B. Write A in its singular value decomposition $A = UDV$, where D is a diagonal matrix with positive values $D_{ii} = d_i$ for $i = 1, \dots p$, and U and V are $p \times p$ orthogonal matrices. Choosing $B = V^{-1}$ gives

$$v_\tau^*(Y) = v_\tau^*(UDZ) = Uv_\tau^*(DZ).$$

Now, by Proposition 4.1, since $d_j \geq 0$ for all j,

$$\mathrm{Var}_\tau(v^\top DZ) = \mathrm{Var}_\tau(\sum_{j=1}^{p} d_j Z_j v_j) = \sum_j v_j^2 d_j^2 \mathrm{Var}_\tau(Z_j).$$

Since $\sum_j v_j^2 = 1$, $\mathrm{Var}_\tau(v^\top DZ)$ lies in the convex hull of the p numbers $d_j^2\mathrm{Var}_\tau(Z_j)$ for $j = 1, \dots p$. Therefore, it is maximized by setting v to be the unit vector along the axis j with maximal $d_j^2\mathrm{Var}_\tau(Z_j)$. But $Z \overset{\mathcal{L}}{=} BZ$ for all orthogonal matrices B, thus $Z_j \overset{\mathcal{L}}{=} Z_k$, hence $\mathrm{Var}_\tau(Z_j) = \mathrm{Var}_\tau(Z_k)$ for all indices $j, k = 1, \dots, p$. Thus $\mathrm{Var}_\tau(v^\top DZ)$ is maximized when v is the unit vector along the axis j with maximal d_j. This is precisely the axis with maximal singular value of A, and hence is also the direction of the (classical) principal component of DZ. This proves the claim. The last statement follows immediately from the third statement. □

We now prove consistency of local maximizers of (7). The main theorem in this section is the following.

Theorem 4.1. *Fix $\tau > 0$. Let Y be a random variable in $\mathbb{R}^p$ with finite second moment, distribution function F. Suppose $v^* = v^*_\tau$ is a unique global solution to (7) corresponding to Y. Suppose we observe n i.i.d copies of Y, with empirical distribution function F_n. Let Y_n be a random variable whose cdf is F_n. Then for sufficiently large n, for any sequence of global solutions v^*_n of (7) corresponding to Y_n, we have*

$$v^*_n \overset{F-a.s.}{\longrightarrow} v^* \text{ in } \mathbb{R}^p \quad \text{as } n \to \infty.$$

For the proof, we first need the following lemma.

Lemma 4.1. *Under the assumptions of Theorem 4.1, uniformly over all $v \in \mathbb{R}^p$ with $v^\top v = 1$, and uniformly over all $\tau \in (0,1)$,*

$$\mathrm{Var}_\tau(Y_n^\top v) \overset{F-a.s.}{\longrightarrow} \mathrm{Var}_\tau(Y^\top v).$$

Proof. Since Y_n is the empirical version of Y and the set of all unit vectors $v \in \mathbb{R}^p, v^\top v = 1$ is compact, by the Cramer-Wold theorem, $Y_n^\top v \xrightarrow{\mathcal{L}} Y^\top v$ uniformly over all such unit vectors $v \in \mathbb{R}^p$. It then follows that e_τ and Var_τ, which are completely determined by the distribution function, also converge $F - a.s.$ uniformly over all v. □

Proof of Theorem 4.1. Let $\mathbb{S}^{p-1}$ denote the unit sphere in $\mathbb{R}^p$. Equip $\mathbb{R}^p$ with the Euclidean norm $\|\cdot\|$. Define the map $V_Y : \mathbb{S}^{p-1} \to \mathbb{R}$, $V_Y(v) = \mathrm{Var}_\tau(Y^\top v)$. Fix $\epsilon > 0$. We shall prove that there exists a $\delta > 0$ such that the global minimum of V_{Y_n} is necessarily within δ-distance of v^*.

Since V_Y is continuous, $\mathbb{S}^{p-1}$ is compact, and v^* is unique, there exists a sufficiently small $\delta > 0$ such that

$$|V_Y(v) - V_Y(v^*)| < \epsilon \Rightarrow \|v - v^*\| < \delta$$

for $v \in \mathbb{S}^{p-1}$. In particular, if $\|v - v^*\| > \delta$, then

$$V_Y(v^*) + \epsilon < V_Y(v).$$

By Lemma 4.1, $V_{Y_n} \to V_Y$ as $n \to \infty$ uniformly over $\mathbb{S}^{p-1}$. In particular, there exists a large N such that for all $n > N$,

$$|V_{Y_n}(v) - V_Y(v)| < \epsilon/6$$

for all $v \in \mathbb{S}^{p-1}$. Thus for $v \in \mathbb{S}^{p-1}$ such that $\|v - v^*\| > \delta$,

$$V_{Y_n}(v) - V_Y(v^*) > \epsilon - \epsilon/6 = 5\epsilon/6.$$

Meanwhile, since V_Y is continuous, one can choose $\epsilon' = \epsilon/6$, and thus obtain δ' such that

$$|V_Y(v) - V_Y(v^*)| < \epsilon/6 \Leftarrow \|v - v^*\| < \delta'.$$

Then, for v such that $\|v - v^*\| < \delta'$,

$$V_{Y_n}(v) - V_Y(v^*) \le |V_{Y_n}(v) - V_Y(v)| + |V_Y(v) - V_Y(v^*)| < \epsilon/6 + \epsilon/6 = \epsilon/3.$$

So far we have shown that if $\|v - v^*\| > \delta$, then $V_{Y_n}(v)$ is at least $5\epsilon/6$ bigger than $V_Y(v^*)$. Meanwhile, if $\|v - v^*\| < \delta'$, then $V_{Y_n}(v)$ is at most $\epsilon/3$ bigger than $V_Y(v^*)$. Thus the global minimum v_n^* of V_{Y_n} necessarily satisfy $\|v_n^* - v^*\| < \delta$. This completes the proof. □

4.1 PEC as constrained PCA

To compute the principal expectile component v_τ^*, one needs to optimize the right-hand side of (7) over all unit vectors v. Although this is a differentiable function in v, optimizing it is a difficult problem, since μ_τ also depends on v, and does not have a closed form solution. However, Theorem 4.2 below shows that in certain situations, for given weights w_i, not only μ_τ but also v_τ^* *have* closed form solutions. In particular, in this setting, PEC is the constrained classical PC of a weighted version of the covariance matrix of the data, centered at a constant possibly different from the mean. This theorem forms the backbone of our iterative algorithm for computing PEC discussed in Section 5.

Theorem 4.2. *Consider (7). Suppose we are given the true weights w_i, which are either τ or $1-\tau$. Let $\tau_+ = \{i \in \{1,\dots,n\} : w_i = \tau\}$ denote the set of observations Y_i with 'positive' labels, and $\tau_- = \{i \in \{1,\dots,n\} : w_i = 1-\tau\}$ denote its complement. Let n_+ and n_- be the sizes of the respective sets. Define an estimator $\hat{e}_\tau \in \mathbb{R}^p$ of the τ-expectile via*

$$\hat{e}_\tau = \frac{\tau \sum_{i\in\tau_+} Y_i + (1-\tau)\sum_{i\in\tau_-} Y_i}{\tau n_+ + (1-\tau)n_-}. \tag{9}$$

Define

$$C_\tau = \frac{\tau}{n}\left\{\sum_{i\in\tau_+}(Y_i - \hat{e}_\tau)(Y_i - \hat{e}_\tau)^\top\right\} + \frac{1-\tau}{n}\left\{\sum_{i\in\tau_-}(Y_i - \hat{e}_\tau)(Y_i - \hat{e}_\tau)^\top\right\}. \tag{10}$$

Then v_τ^* is the solution to the following optimization problem:

$$\begin{aligned} &\text{maximize } v^\top C_\tau v \\ &\text{subject to } v^\top Y_i > v^\top \hat{e}_\tau \Leftrightarrow i \in \tau_+ \\ &\qquad\qquad v^\top v = 1. \end{aligned} \tag{11}$$

Proof. Since the weights are the true weights coming from the true principal expectile component v_τ^*, clearly v_τ^* satisfies the constraint in (11). Now suppose v is another vector in this constraint set. Then $v^\top \hat{e}_\tau$ is exactly μ_τ, the τ-expectile of the sequence of n real numbers $v^\top Y_1, \dots, v^\top Y_n$. Therefore, the quantity we need to maximize in (7) reads

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n (v^\top Y_i - \mu_\tau)^2 w_i &= \frac{\tau}{n}\sum_{i\in\tau_+}(v^\top Y_i - v^\top \hat{e}_\tau)^2 + \frac{1-\tau}{n}\sum_{i\in\tau_-}(v^\top Y_i - v^\top \hat{e}_\tau)^2 \\ &= \frac{\tau}{n}\sum_{i\in\tau_+} v^\top (Y_i - \hat{e}_\tau)(Y_i - \hat{e}_\tau)^\top v + \frac{1-\tau}{n}\sum_{i\in\tau_-} v^\top (Y_i - \hat{e}_\tau)(Y_i - \hat{e}_\tau)^\top v \\ &= v^\top C_\tau v. \end{aligned}$$

Thus the optimization problem above is indeed an equivalent formulation of (7), which was used to define v_τ^*. Finally, the last observation follows by comparing the above with the optimization formulation for PCA, see the paragraph after (4). Indeed, when $\tau = 1/2$, $\hat{e}_{1/2} = \bar{Y}$, $C_{1/2} = C$, and we recover the classical PCA. □

5 Algorithms

5.1 Principal Expectile Components

Suppose the conditions of Theorem 4.2 are satisfied, so finding PEC is the problem of solving a constrained PCA given in (11), but with unknown weights depending on the true principal direction. Since $\hat{e}_\tau$ is a linear function in the Y_i, (11) defines a system of linear constraints in the entries of Y_i and v_τ^*. Thus for each fixed sign sets (τ_+, τ_-), there exist (not necessarily unique) local optima $v_\tau^*(\tau_+, \tau_-)$. There are 2^n possible sign sets, one of which corresponds to the global optima v_τ^* that we need. It is clear that finding the global optimum v_τ^* by enumerating all possible sign sets is intractable. However, in many situations, the constraint in (11) is inactive. That is, the largest eigenvector of C_τ satisfies (11) for free. In such situations, we call v_τ^* a *stable solution*. Just like classical PCA, stable solutions are unique for matrices C_τ with unique principal eigenvalue. More importantly, we have an efficient algorithm for finding stable solutions, if they exist.

Definition 5.1. For some given sets of weights $w = (w_i)$, define $e_\tau(w)$ via (9), $C_\tau(w)$ via (10). Let $v_\tau(w)$ be the largest eigenvector of $C_\tau(w)$. If $v_\tau(w)$ satisfies (11), we say that $v_\tau(w)$ is a locally *stable solution* with weights w.

To find locally stable solutions, one can solve (3) using iterative reweighted least squares: first initialize the w_i's, compute estimators $\mu_\tau(w)$ and $v_\tau(w)$ ignoring the constraint (11), update the weights via (8), and iterate. At each step of this algorithm, one finds the principal component of a weighted covariance matrix with some approximate weight. Since there are only finitely many possible weight sets, the algorithm is guaranteed to converge to a locally stable solution if it exists. In particular, if the true solution to (3) is stable, then for appropriate initial weights, the algorithm will find this value. We call this algorithm *PrincipalExpectile*.

We now describe the details of this algorithm for the case $k = 1$, that is, the algorithm for computing the first principal expectile component only. To obtain higher order components, one iterates the algorithm over the residuals $Y_i - \hat{v}_1(\hat{v}_1^\top Y_i + \hat{\mu}_1)$, where $\hat{\mu}_1$ is the τ-expectile of the loadings $\hat{v}_1^\top Y_i$.

For n observations $Y_1, \ldots, Y_n$, there are at most 2^n possible labels for the Y_i's, and hence the algorithm has in total 2^n possible values for the w_i's. Thus either Algorithm 1 converges to a point which satisfies the properties of the optimal solution that Theorem 4.2 prescribes, or that it iterates infinitely over a cycle of finitely many possible values of the w_i's. In particular, the true solution is a fixed point, and thus fixed points always exist. In practice, we find that the algorithm converges very quickly, and can get stuck in a finite cycle of values. In this case, one can jump to a different starting point and restart the algorithm. Choosing a good starting value is important in ensuring convergence. Since the τ-variance is a continuous function in τ, we find that in most cases, one can choose a good starting point by performing a sequence of

Algorithm 1 PrincipalExpectile

1: Input: data $Y \in \mathbb{R}^{n \times p}$.
2: Output: a vector $\hat{v}$, an estimator of the first principal expectile component of Y.
3: **procedure** PRINCIPALEXPECTILE(Y)
4: Initialize the weights $w_i^{(0)}$
5: Set $t = 0$.
6: **repeat**
7: Let $\tau_+^{(t)}$ be the set of indices i such that $w_i^{(t)} = \tau$, and $\tau_-^{(t)}$ be the complement.
8: Compute $e_\tau^{(t)}$ as in equation (9) with sets $\tau_+^{(t)}, \tau_-^{(t)}$.
9: Compute $C_\tau^{(t)}$ as in equation (10) with sets $\tau_+^{(t)}, \tau_-^{(t)}$.
10: Set $v^{(t)}$ to be the largest eigenvector of $C_\tau^t (C_\tau^t)^\top$
11: Set $\mu_\tau^{(t)}$ to be the τ-expectile of $(v^{(t)})^\top Y_i$
12: Update w_i: set $w_i^{(t+1)} = \tau$ if $(v^{(t)})^\top Y_i > \mu_\tau^{(t)}$, and set $w_i^{(t+1)} = 1 - \tau$ otherwise.
13: Set t = t + 1
14: **until** $w_i^t = w_i^{(t+1)}$ for all i.
15: **return** $\hat{v} = v^{(t)}$.
16: **end procedure**

such computations for a sequence of τ starting with $\tau = 1/2$, and set the initial weight to be that induced by the previous run of the algorithm for a slightly smaller (or larger) τ.

5.2 TopDown and BottomUp

We now describe how iterative weighted least squares can be adapted to implement TopDown and BottomUp. We start with a description of the asymmetric weighted least squares (LAWS) algorithm of Newey and Powell (1987). The basic algorithm outputs a subspace without the affine term, and needs to be adapted. See Guo et al. (2015) for a variation with smoothing penalty and spline basis.

Proposition 5.1. *The LAWS algorithm is well-defined, and is a gradient descent algorithm. Thus it converges to a critical point of the optimization problem* (1).

Proof. First, we note that the steps in the algorithm are well-defined. For fixed W and V, $J(U, V, W)$ is a quadratic in the entries of U. Thus the global minimum on line 8 has an explicit solution, see Srebro and Jaakkola (2003); Guo et al. (2015). A similar statement applies to line 9.

Note that $J(U, V, W)$ is not jointly convex in U and V, but as a function in U for fixed V, it is a convex, continuously differentiable, piecewise quadratic function. The statement holds for $J(U, V, W)$ as a function in V for fixed U. Hence lines 8 and 9 is one step in a Newton-Raphson algorithm on $J(U, V, W)$ for fixed V. Similarly, lines 10 and 11 is one step in a Newton-Raphson algorithm on $J(U, V, W)$ for fixed U. Thus the algorithm is a coordinatewise gradient descent on a coordinatewise convex function, hence converges. □

If some columns of U or V are pre-specified, one can run LAWS and not update these columns in lines 8 and 10. Thus one can use LAWS to find the optimal affine subspace by writing $\mathbf{1}m^\top + E = \tilde{U}\tilde{V}$ with the first column of $\tilde{U}$ constrained to be $\mathbf{1}$. Similarly, we can use this technique to solve the constrained optimization problems:

Algorithm 2 Asymmetric weighted least squares (LAWS)

Input: data $Y \in \mathbb{R}^{n\times p}$, positive integer $k < p$
Output: $\hat{E}_k^*$, an estimator of E_k^*, expressed in product form $\hat{E}_k^* = \hat{U}\hat{V}^\top$, where $\hat{U} \in \mathbb{R}^{n\times k}, \hat{V} \in \mathbb{R}^{p\times k}.\hat{U}, \hat{V}$ are unique up to multiplication by an invertible matrix.
procedure LAWS(Y, k)
    Set $V^{(0)}$ to be some rank-k $p \times k$ matrix.
    Set $W^{(0)} \in \mathbb{R}^{n\times p}$ to be $1/2$ everywhere.
    Set $t = 0$.
    repeat
        Update U: Set $U^{(t+1)} = \text{argmin}_{U\in\mathbb{R}^{n\times k}} J(U, V^{(t)}, W^{(t)})$.
        Update W: Set $W_{ij}^{(t+1)} = \tau$ if $Y_{ij} - \sum_l U_{il}^{(t+1)} V_{lk}^{(t)} > 0$, $W_{ij}^{(t+1)} = 1 - \tau$ otherwise.
        Update V: Set $V^{(t+1)} = \text{argmin}_{V\in\mathbb{R}^{k\times p}} J(U^{(t+1)}, V, W^{(t+1)})$.
        Update W: Set $W_{ij}^{(t+1)} = \tau$ if $Y_{ij} - \sum_l U_{il}^{(t+1)} V_{lk}^{(t+1)} > 0$, $W_{ij}^{(t+1)} = 1 - \tau$ otherwise.
        Set t = t + 1
    until $U^{(t+1)} = U^{(t)}, V^{(t+1)} = V^{(t)}, W^{(t+1)} = W^{(t)}$.
return $\hat{E}_k = U^{(t)}(V^{(t)})^\top$.
end procedure

- *Find a rank-k approximation E_k whose span contains a given subspace of dimension $r < k$.*
- Solution: Constrain the first r columns of $V^{(0)}$ to be a basis of the given subspace.
- *Find a rank-k approximation whose span lies within a given subspace of dimension $r > k$.*
- Solution: Let $B \in \mathbb{R}^{n\times r}$ be a basis of the given subspace. Then the optimization problem becomes

$$\min_{U\in\mathbb{R}^{r\times k}, V\in\mathbb{R}^{p\times k}} \|Y - BUV^\top\|_{\tau,2}^2.$$

  One can then apply the LAWS algorithm with variables U and V.
- *Find a rank-k approximation whose span contains a given subspace of dimension $r < k$, and is contained in a given subspace of dimension $R > k$.*
- Solution: Combine the previous two solutions.

Algorithm 3 TopDown

Input: data $Y \in \mathbb{R}^{n\times p}$, positive integer $k < p$
Output: $\hat{E}_k^*$, an estimator of E_k^*, expressed in product form $\hat{E}_k^* = \hat{U}\hat{V}^\top$, where $\hat{U} \in \mathbb{R}^{n\times k}, \hat{V} \in \mathbb{R}^{p\times k}$ are unique.
procedure TopDown(Y, k)
    Use LAWS(Y,k) to find $\hat{m}_k^*, \hat{E}_k^*$. Write $\hat{E}_k^* = UV^\top$ for some orthonormal basis U.
    Use LAWS to find $\hat{U}_1$, the vector which spans the optimal subspace of dimension 1 contained in U.
    Use LAWS to find $\hat{U}_2$, where $(\hat{U}_1, \hat{U}_2)$ spans the optimal subspace of dimension 1 contained in U and contains the span of $\hat{U}_1$
    Repeat the above step until obtains $\hat{U}$.
    Obtain $\hat{V}$ through the constraint $\hat{E}_k^* = \hat{U}\hat{V}^\top$.
return $\hat{m}_k^*, \hat{E}_k^*, \hat{U}, \hat{V}^\top$.
end procedure

Algorithm 4 BottomUp

Input: data $Y \in \mathbb{R}^{n\times p}$, positive integer $k < p$
Output: $\hat{E}_k^*$, an estimator of E_k^*, expressed in product form $\hat{E}_k^* = \hat{U}\hat{V}^\top$, where $\hat{U} \in \mathbb{R}^{n\times k}, \hat{V} \in \mathbb{R}^{p\times k}$ are unique.
procedure BottomUp(Y, k)
    Use LAWS to find $\hat{E}_1^*$. Let $\hat{U}_1$ be the basis vector.
    Use LAWS to find $\hat{U}_2$ such that $(\hat{U}_1, \hat{U}_2)$ is the best two-dimensional approximation to Y, subjected to containing $\hat{U}_1$.
    Repeat the above step until obtains $\hat{U}$. We obtain $\hat{V}$ and $\hat{E}_k^*$ in the last iteration. **return** $\hat{E}_k^*, \hat{U}, \hat{V}^\top$.
end procedure

With these tools, we now define the two algorithms, *TopDown* and *BottomUp*. The TopDown algorithm requires the weights w_{ij} and the loadings on previous principal components to be re-evaluated when finding the next principal component. A variant of the algorithm would be to keep the weights w_{ij}. In this case, the algorithm is still well-defined. However, it will produce a different basis matrix $\hat{U}$, since the estimators are no longer optimal in the $\|\cdot\|_{\tau,2}^2$ norm.

5.3 Performance bounds of TopDown and BottomUp

We now show that the dependence on k only grows polylog in n. Thus both TopDown and BottomUp are fairly efficient algorithms even for large k.

Theorem 5.1. *For fixed V of dimension k, LAWS requires at most $\mathcal{O}\{\log(p)^k\}$ iterations, $\mathcal{O}\{npk^2\log(p)^k\}$ flops to estimate U.*

In other words, if V has converged, LAWS needs at most $\mathcal{O}\{npk^2\log(p)^k\}$ flops to estimate U. The role of U and V are interchangeable if we transpose Y. Thus if U has converged, LAWS needs at most $\mathcal{O}\{npk^2\log(n)^k\}$ to estimate V. We do not have a bound for the number of iterations needed until convergence. In practice this seem to be of order log of n and p. For the proof of Theorem 5.1 we need the following two lemmas.

Lemma 5.1. *If $Y_1, \ldots, Y_n \in \mathbb{R}$ are n real numbers, then LAWS finds their τ-expectile e_τ in $\mathcal{O}\{\log(n)\}$ iterations.*

Proof. Given the weights $w_1, \ldots, w_n$, that is, given which Y_i's are above and below e_τ, the τ-expectile e_τ is a linear function in the Y_i as we saw in (9). As shown in Proposition 5.1, LAWS is equivalent to a Newton-Raphson algorithm on a piecewise quadratic function. Since the points Y_i's are ordered, it takes $\mathcal{O}\{\log(n)\}$ to learn their true weights. Thus the algorithm converges in $\mathcal{O}\{\log(n)\}$ iterations. □

Lemma 5.2. *An affine line in $\mathbb{R}^p$ can intersect at most $2p$ orthants.*

Proof. Recall that an *orthant* of $\mathbb{R}^p$ is a subset of $\mathbb{R}^p$ where the sign of each coordinate is constrained to be either nonnegative or nonpositive. There are 2^p orthants in $\mathbb{R}^p$. Let $f(\lambda) = Y + \lambda v$ be our affine line, $\lambda \in \mathbb{R}, Y, v \in \mathbb{R}^p$. Let $\text{sgn} : \mathbb{R}^p \to \{\pm 1\}^p$ denote the sign function. Now, $\text{sgn}\{f(0)\} = \text{sgn}(Y), \text{sgn}\{f(\infty)\} = \text{sgn}(v)$, and $\text{sgn}\{f(\lambda)\}$ is a monotone increasing function in λ. As $\lambda \to \infty$, $\text{sgn}\{f(\lambda)\}$ goes from $\text{sgn}(Y)$ to $\text{sgn}(v)$ one bit flip at a time. Thus there are at most p flips, that is, the half-line $f(\lambda)$ for $\lambda \in [0, \infty)$ intersects at most p orthants. By a similar argument, the half-line $f(\lambda)$ for $\lambda \in (-\infty, 0)$ intersects at most p other orthants. This concludes the proof. □

Corollary 5.1. *An affine subspace of dimension k in $\mathbb{R}^p$ can intersect at most $\mathcal{O}(p^k)$ orthants.*

Proof. Fix any basis, say $\psi_1, \ldots, \psi_k$. By Lemma 5.2, ψ_1 can intersect at most $2p$ orthants. For each orthant of ψ_1, varying along ψ_2 can yield at most another $2p$ orthants. The proof follows by induction. (This is a rather liberal bound, but it is of the correct order for k small relative to p). □

Proof of Theorem 5.1. By Corollary 5.1, it is sufficient to consider the case $k = 1$. Fix V of dimension 1. Since U, V are column matrices, we write them in lower case letters u, v. Solving for each u_i is a separate problem, thus we have n separate optimization problem, and it is sufficient to prove the claim for each i for $i = 1, \ldots, n$.
Fix an i. As u_i varies, $Y_i - m_i - u_i v$ defines a line in $\mathbb{R}^p$. The weight vector $(w_{i1}, \ldots, w_{ip})$ only depend on which coordinates are the orthant of $\mathbb{R}^p$ in which $Y_i - m_i - u_i v$ is in. The later is equivalently to determining the weight of the p points $\frac{Y_i - m_i}{v_i}$. By Lemma 5.1, it takes $\mathcal{O}\{\log(p)\}$ for LAWS to determine the weights correctly. Thus LAWS takes at most $\mathcal{O}\{\log(p)\}$ iterations to converge, since each iteration involves estimating w, then v. Each iteration solves a weighted least squares, thus take $\mathcal{O}(npk^2)$. Hence for fixed v, LAWS can estimate u after at most $\mathcal{O}\{npk^2 \log(p)\}$ flops for $k = 1$. This concludes the proof for fixed v. By considering the transposed matrix Y, we see that the role of u and v are interchangeable. The conclusion follows similarly for fixed u. □

6 Simulation

To study the finite sample properties of the proposed algorithms we do a simulation study. We follow the simulation setup of Guo et al. (2015), that is, we simulate the data $Y_{ij}, i = 1, \ldots, n$, $j = 1, \ldots, p$ as

$$Y_{ij} = \mu(t_j) + f_1(t_j)\alpha_{1i} + f_2(t_j)\alpha_{2i} + \varepsilon_{ij}, \tag{12}$$

where t_j's are equidistant on [0,1], $\mu(t) = 1 + t + \exp\{-(t-0.6)^2/0.05\}$ is the mean function, $f_1(t) = \sqrt{2}\sin(2\pi t)$ and $f_2(t) = \sqrt{2}\cos(2\pi t)$ are principal component curves, and ε_{ij} is a random noise.
We consider different settings 1 and 2 each with five error scenarios:

1. $\alpha_{1i} \sim \mathrm{N}(0, 36)$ and $\alpha_{2i} \sim \mathrm{N}(0, 9)$ are both iid and ε_{ij}'s are (1) iid $\mathrm{N}(0, \sigma_1^2)$, (2) iid $t(5)$, (3) independent $\mathrm{N}\{0, \mu(t_j)\sigma_1^2\}$, (4) iid $\mathrm{logN}(0, \sigma_1^2)$ and (5) iid sums of two uniforms $U(0, \sigma_1^2)$ with σ_1^2=0.5.

2. $\alpha_{1i} \sim \mathrm{N}(0, 16)$ and $\alpha_{2i} \sim \mathrm{N}(0, 9)$ are both iid and ε_{ij}'s are (1) iid $\mathrm{N}(0, \sigma_2^2)$, (2) iid $t(5)$, (3) independent $\mathrm{N}\{0, \mu(t_j)\sigma_2^2\}$, (4) iid $\mathrm{logN}(0, \sigma_2^2)$ and (5) iid sums of two uniforms $U(0, \sigma_2^2)$ with σ_2^2=1.

Note that the settings imply different ratios of coefficient-to-coefficient-to-noise variations. In the setting 1 scenario (1) we have a ratio 36:9:0.5, whereas in the setting 2 scenario (1) we have 16:9:1. Apart from standard Gaussian errors, we also consider "fat tailed" errors in scenario (2), heteroscedastic in (3) and skewed errors in (4). We study the performance of the algorithms for three sample sizes: (i) small n=20, p=100; (ii) medium n=50, p=150; (iii) large n=100, p=200.

For every combination of parameters we repeat the simulations 500 times and record the mean computing times, the mean of the average mean squared error (MSE), its standard deviation, and convergence ratio for each algorithm. We label the run of the algorithm as unconverged whenever after 30 iterations and 50 restarts from a random starting point the algorithms fail to converge.

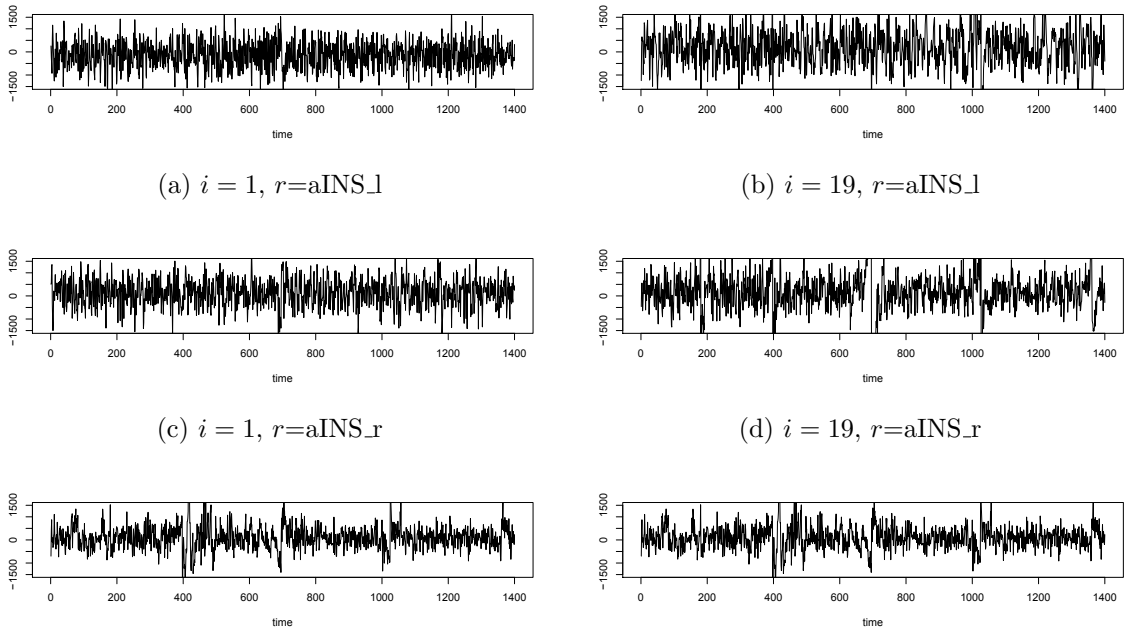
We compare computational times and MSEs of the three methods TopDown (TD), BottomUp (BUP) and PrincipalExpectile (PEC) in the Appendix. In general, PEC is the fastest, however, it has lower convergence rate than TopDown (TD) and BottomUp (BUP). From the MSEs, we conclude that whenever the error distribution is fat-tailed or skewed, or by small samples PEC is likely to produce more reliable results in terms of its MSE, whereas by errors close to normal and moderate or large samples TD is likely to produce smaller MSEs.

7 Empirical Study

We apply the proposed algorithms to two different datasets. In section 7.1 we investigate the fMRI data from Risk Perception in Investment Decisions (RPID) study. Since the technical details of experiment are complex and beyond the scope of this research, we provide the extended extended introductory summary of experiment and refer the reader to Mohr et al. (2010), Mohr and Nagel (2010) and Majer et al. (2015) for more details about experiment or book of Ashby (2011) for analysis fMRI data in general. In section 7.2 we analyze the daily temperature dataset over multiple Chinese stations.

7.1 Application to fMRI data

Risk Perception in Investment Decisions (RPID) Study performed an experiment over 19 individuals. Each participant was asked 256 investment questions, where past returns were presented and participants had to make a choice whether they would invest in a bond with 5% fixed return or the displayed investment. Individual responses reflect the risk attitude of every participant.



(a) $i = 1$, r=aINS_l (b) $i = 19$, r=aINS_l

(c) $i = 1$, r=aINS_r (d) $i = 19$, r=aINS_r

(e) $i = 1$, r=DMPFC (f) $i = 19$, r=DMPFC

Figure 1: Loadings for the 1-st principal expectile component for active regions of individual No.1. (left) and No. 19 (right).

PEC_fmri

Following the common Markowitz mean-variance approach one can evaluate the this risk attitude, see details Mohr and Nagel (2010) and assign the corresponding values between -0.1 and 1.1 reflecting individual risk perception. We show the values in Figure 2 (right) on vertical axes. Higher values represent the higher risk aversion. The individual No.19 is considered as the most risk seeking and individual No.1 as the most risk averse participant in the population sample.

The aim of experiment was to study if individual's risk perception can be interpreted and recovered by brain activities. With Functional magnetic resonance imaging (fMRI) one can measure such neural activity by the blood oxygen level-dependent (BOLD) signal.

Regarding the settings of our dataset, scans of voxels were taken every 2 seconds and as a result the high-dimensional data were obtained for each individual. Majer et al. (2015) identified three brain regions (clusters) which are activated during the experiment: anterior insula (aINS; left and right) and dorsomedial prefrontal cortex (DMPFC). From a statistical point of view the scan of all voxels in certain brain area can be considered as a multi-dimensional time series of round 300-400 voxels for every individual, however very noisy. In order to capture the variability in these series of every region we use principal expectile components.

Following the notation from Section 4, denote $Y_t^{(r)}$ the response, $N^{(r)}$-dimensional vector, obtained at specific region, $r = aINS_{left}, aINS_{right}, DMPFC$ at time $t = 1, \ldots, 1400$, where $N^{(r)}$ is a number of voxels in a specific region r. Further, $\phi^k_{\tau,(r)}$ its k-th principal expectile component (PEC) at level τ and $\psi^k_{\tau,(r)}$ corresponding projections, also known as loadings. PECs provide us with necessary dimension reduction; each region dynamics is now captured by univariate time series of loadings. The loadings of all three regions for indiviuals No.1 and No.19 at level $\tau = 0.6$ are presented in Figure 1.

Since the response function usually achieves the peak only shortly after stimulus, i.e. portfolio question, we focus on average loadings after stimulus. The average loadings of three active regions are considered as the regressors for explanation of this risk attitude. In order to be able to compare the results of previous work, we follow Majer et al. (2015) and model the relationship of the risk attitude att_i and brain reactions via linear regression, which provides the simplest however quite accurate comparison. More precisely for individuals $i = 1, \ldots, 19$ and any fixed τ-level we have:

$$att_i = \beta_0 + \sum_{\substack{k=1,2 \\ r=aINS\_l, \\ aINS\_r, \\ DMPFC}} \beta_{k,r} \overline{\psi^k_{\tau,(r),i}} + \varepsilon_i. \tag{13}$$

We performed the PrincipalExpectile algorithm for different $\tau = 0.05, 0.1, \ldots, 0.9, 0.95$. It is interesting to see that the best result with respect to coefficient of determination R^2 is obtained not for $\tau = 0.5$, however for $\tau = 0.6$. We report the coefficients of determinations for all considered τ-level in Figure 2 together with the regression fit for model (13) at level $\tau = 0.6$. The usage of $\tau = 0.6$ over-performs the traditional usage of $\tau = 0.5$.

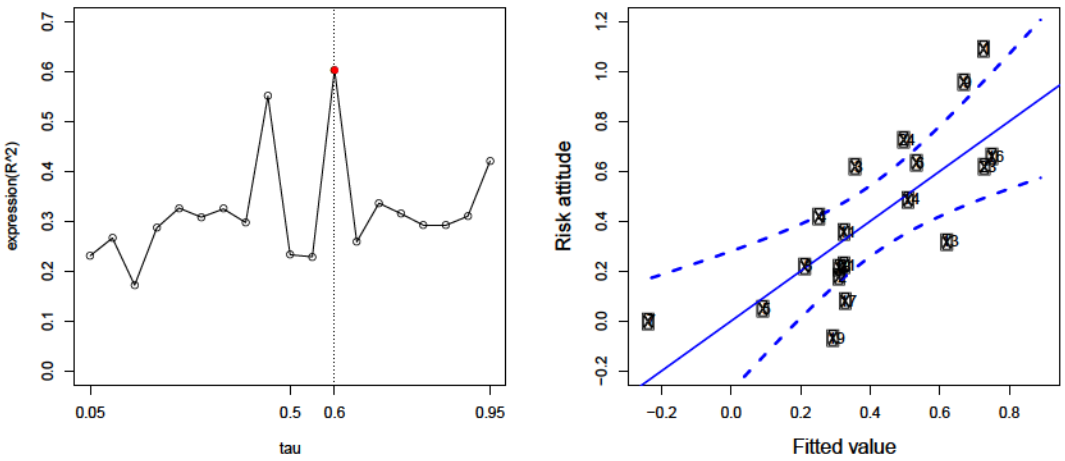


Figure 2: Coefficients of determinations for all considered τ-level (left) and the regression fit for model (13) at $\tau = 0.6$ (right). Horizontal axis represents $\widehat{att_i}$, the best linear combination of regressors $\overline{\psi^k_{\tau,(r),i}}$.

PEC_fmri

7.2 Application to Chinese Weather Data

We apply the algorithms BottomUp, TopDown and PrincipalExpectile to Chinese temperature data using daily average temperature data of 159 weather stations in mainland China for the years 1957 to 2009 provided by Chinese Meteorological Administration via its website. We did not pre-smooth the data. Original data averaged over years for every country are presented in Figure 3.

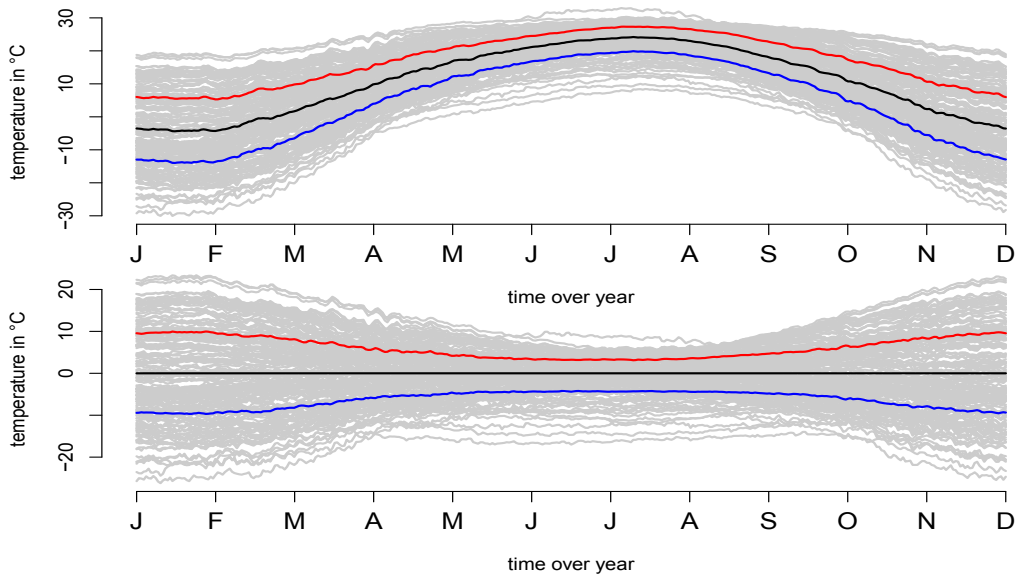


Figure 3: Observed averaged daily temperature on 159 stations (upper panel) and decentred data (lower panel) with expectiles for level τ= 0.9, 0.5 and 0.1.

PEC_temperature

We run the algorithms to estimate principal expectile components for the weather stations at each of the τ-levels 10%, 50% and 90% with respect to days of a year from 1 to 365. Our analysis for the 50% expectile corresponds to the classical PCA. We estimate first two principal component functions.The estimation results of the three proposed algorithms are rather similar. In Figure 4 we present the estimated principal component functions for $\tau = 0.1$ and $\tau = 0.9$.

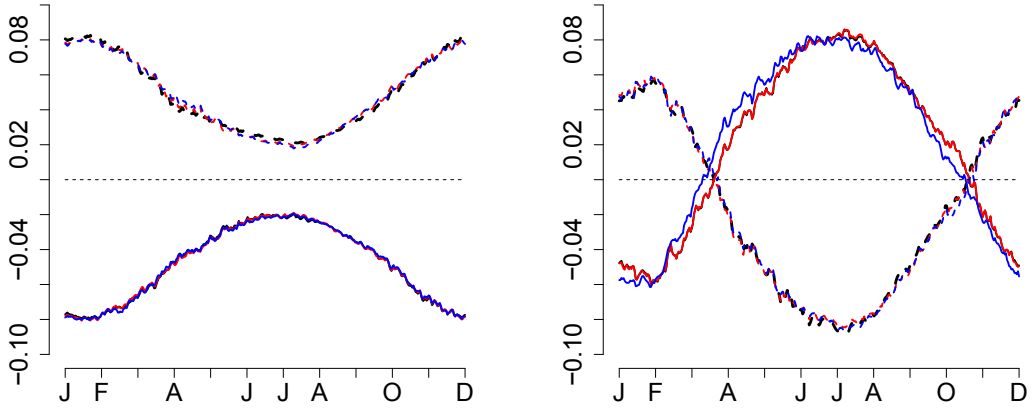


Figure 4: The estimated first PEC (left) and 2nd PEC (right) for $\tau = 0.1$ (dashed) and $\tau = 0.9$ (solid, multiplied by -1) computed with three proposed algorithms TopDown (red), BottomUp (black) and PrincipileExpectile (blue).

PEC_temperature

We see that all three algorithms give really similar results. However, one can be more interested in differences among the levels of τ. Thus, in Figure 5 we show the differences of PEC component at level $\tau = 0.9$ (red), $\tau = 0.1$ (blue) respectively, and PEC component at level $\tau = 0.5$, which corresponds to the ordinary principal component, i.e. $\phi^k_{0.9} - \phi^k_{0.5}$, resp. $\phi^k_{0.1} - \phi^k_{0.5}$. We observe that both components differ from ordinary principal compoment. Moreover, we plot also differences for $\tau = 0.8$ (dashed gray) and $\tau = 0.7$ (solid gray) to show that in case of the 2nd component, the difference increases with higher level of τ.

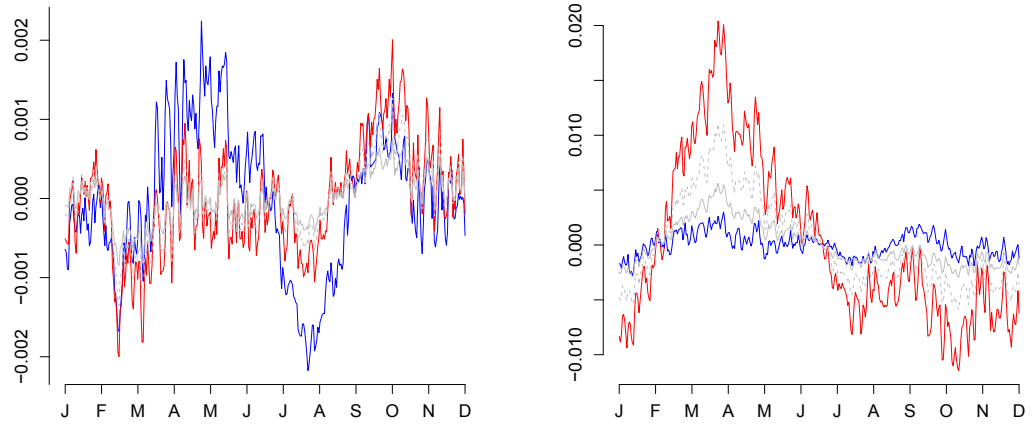


Figure 5: The differences of estimated PECs for $\tau = 0.1$ (blue) and $\tau = 0.9$ (red, multiplied by -1) from estimeted PEC for $\tau = 0.5$, computed with PrincipileExpectile algorithm. Differences for 1st component are shown in left, for 2nd component in right.

PEC_temperature

The obtained first and second components indicate the changes in the temperature distribution from lighter to heavier tails and the other way around within a typical year. A positive score on the first component would mean heavier than average tails in winter and lighter than average tails during the rest of the year. Similar, a positive score on the second component would indicate lighter than average tails of the temperature distribution in winter months, and heavier in summer.

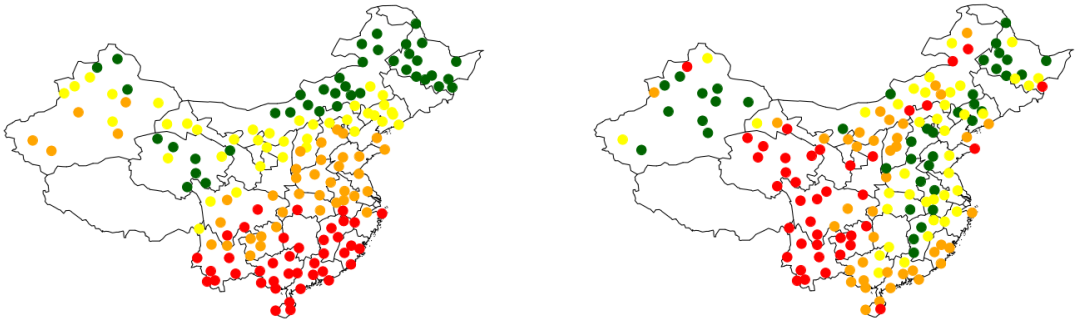


Figure 6: The scores for 1st (left) and 2nd (right) PECs computed by Principile expectile algorithm for $\tau = 0.9$.

PEC_temperature

.

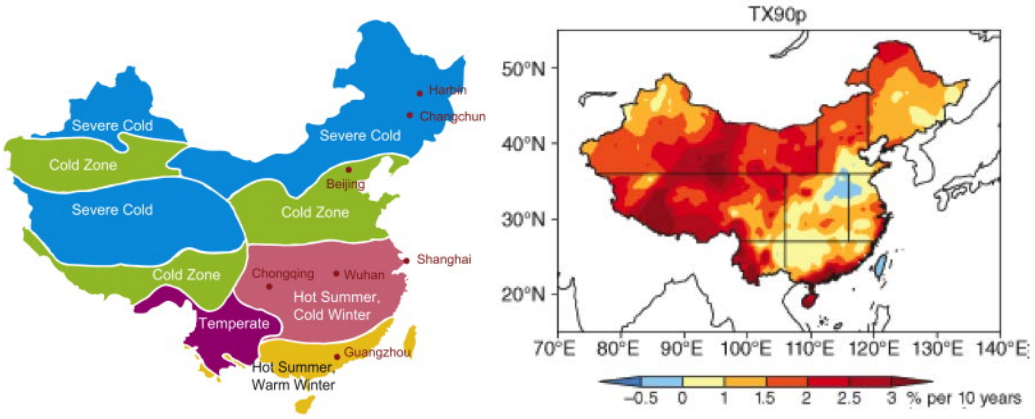


Figure 7: Map of Chinese climate zones by Gao et al. (2014) (left) and distribution of trends in temperature percentile index TX90p for the period 1961-2010 by Zhou et al. (2016) (right).

For better interpretation we also provide the map of scores in Figure 6. The scores of the first principal expectile component correspond to the climate regions, see Figure 7, explaining the short-term periodic behaviour precisely. On the contrary, the scores of the second principal expectile component correspond more to the increasing trend in extremes observed for different areas, shown in Figure 7 via temperature index TX90p. TX90p - Warm days indicator is a percentage of time when daily max temperature is higher than 90th percentile. It is also of the 27 core indicators for temperature and perspiration recommended by WMO - ETCCDI (World Meteorological Organization- Expert Team on Climate Change Detection and Indices),

see Klein Tank et al. (2009). It is obvious that the scores of second component do not necessary coincide with the climate regions but with areas of TX90p index, which explains more long-term behaviour and trend.

8 Summary

We proposed two definitions of principal components in an asymmetric norm and provided consistent algorithms based on iterative least squares. We derived the upper bounds on their convergence times as well as other useful properties of the resulting principal components in an asymmetric norm.

The algorithms TopDown and BottomUp minimize the projection error in a τ-asymmetric norm, and PrincipalExpectile algorithm maximizes the τ-variance of the low-dimensional projection. The later algorithm was shown to share 'nice' properties of PCA as invariance under translations and changes of basis, moreover, it coincides with classical PCA for elliptically symmetric distributions. In simulations, PrincipalExpectile and TopDown have very satisfactory performances in terms of the MSE. In addition, PrincipalExpectile showed robustness to 'fat-tails' and skewness of the data distribution.

We applied the algorith to fMRI data to analyze the possibility of better explanation of individual risk attitude by brain reactions. We have shown that one can achieve a better results with a help of higher τ-level rather than by commonly used $\tau = 0.5$.

We also applied the algorithms to Chinese weather dataset with a view to analyzing weather extremes and long-term behaviour. Analogously to principal components by Ramsay and Silverman (2005) we estimated the first two principal expectile component functions of the temperature as functions of days over a year. The resulting component functions indicate relative changes in the tails of the temperature distribution from light to heavier and vice versa. Our further results clarify the meaning of 1st component as seasonal component explaining short-term variance of climate areas, while the 2nd component corresponds to the long-term changes.

The proposed algorithms appear to be a good way to study extremes of multivariate data. They are easy to compute, relatively fast and their results are easy to interpret.

9 Appendix

Table 1 and 2 show the runtimes of the simulations. PrincipalExpectile (PEC) is the fastest algorithm, however, it has relative low convergence rate: for all sample sizes only around 80% of algorithm runs were convergent. In 20% cases the algorithm keeps iterating between two sets of weights which possibly indicates an adverse sample geometry, i.e. that two eigevalues of the scaled covariance matrix are too close to each other. TD, on the contrary, converges almost always in medium and large sample sizes.

sample	small			medium			large		
τ/sec	BUP	TD	PEC	BUP	TD	PEC	BUP	TD	PEC
0.900	1.15	0.70	0.57	2.87	1.59	1.39	7.44	4.02	2.71
0.950	1.52	1.13	0.55	3.94	2.68	1.57	10.34	6.88	3.03
0.975	2.47	2.32	0.56	5.49	4.62	1.56	14.37	10.96	3.54

Table 1: Average time in seconds for convergence of the algorithms by 500 simulations

sample	small			medium			large		
τ/rate	BUP	TD	PEC	BUP	TD	PEC	BUP	TD	PEC
0.900	0.11	0.00	0.24	0.07	0.00	0.23	0.03	0.00	0.20
0.950	0.17	0.00	0.22	0.13	0.00	0.26	0.11	0.00	0.21
0.975	0.25	0.03	0.21	0.22	0.01	0.25	0.22	0.00	0.24

Table 2: Nonconvergence rates of the algorithms by 500 simulation runs

The results on the MSEs for both simulation settings are presented in Tables 3 and 4. For the settings 1 and 2 solely the magnitude of the average MSE differs; there is no substantial qualitative difference in relative performance of the algorithms. BUP performs the worst of the three algorithms in terms of its MSE in all scenarios. TD and PEC are comparable in terms of their MSEs. PEC shows robustness against skewness and fat tails in the error distribution since it produces the lowest MSEs in scenarios (2) and (4). Yet TD tends to slightly outperform PEC in medium and large samples by errors close to iid normal or normal heteroscedastic; by small sample sizes PEC outperforms TD in all scenarios but (5).

Figures 8 and 9 illustrate the difference in the quality of component estimation for the 95% expectile when coefficient-to-coefficient-to-noise variation ratio changes (setting 1 versus setting 2 respectively). The results are shown for the error scenario (1) and small sample size. We observe that as the ratio changes from 36:9:0.5 (setting 1, Figure 8) to 16:9:1 (setting 2, Figure 9) the variability of the estimators of both component functions increases. The overall mean of the estimators remains very close to the true component functions.

scenario	τ	n=20, p=100			n=50, p=150			n=100, p=200		
		BUP	TD	PEC	BUP	TD	PEC	BUP	TD	PEC
(1)	0.900	0.2762	0.1216	0.1123	0.1339	0.0538	0.0632	0.0698	0.0297	0.0459
		(0.1997)	(0.0097)	(0.0111)	(0.1099)	(0.0033)	(0.0029)	(0.0552)	(0.0015)	(0.0014)
	0.950	0.3619	0.1568	0.1334	0.2323	0.0705	0.0727	0.1312	0.0394	0.051
		(0.2199)	(0.0123)	(0.0181)	(0.2076)	(0.0045)	(0.0044)	(0.1415)	(0.0020)	(0.0019)
	0.975	0.5064	0.2053	0.1601	0.3583	0.0944	0.0874	0.2157	0.0536	0.0594
		(0.2977)	(0.0154)	(0.0276)	(0.2989)	(0.0060)	(0.0075)	(0.2314)	(0.0027)	(0.0035)
(2)	0.900	0.7092	0.5421	0.3147	0.3382	0.2714	0.1494	0.1866	0.1548	0.0932
		(0.2382)	(0.1096)	(0.0685)	(0.1223)	(0.0727)	(0.0117)	(0.0522)	(0.0217)	(0.0050)
	0.950	1.105	0.7847	0.3854	0.5789	0.4440	0.1819	0.3316	0.2680	0.1101
		(0.4453)	(0.1646)	(0.0988)	(0.2664)	(0.1675)	(0.0192)	(0.1144)	(0.0575)	(0.0075)
	0.975	1.6066	1.1158	0.4709	0.9956	0.7033	0.2309	0.5780	0.4641	0.1358
		(0.7968)	(0.2106)	(0.1413)	(0.6936)	(0.2629)	(0.0341)	(0.2227)	(0.1175)	(0.0132)
(3)	0.900	0.4146	0.2300	0.2215	0.1829	0.1019	0.1270	0.0962	0.0562	0.0942
		(0.2413)	(0.0195)	(0.0236)	(0.1070)	(0.0065)	(0.0066)	(0.0510)	(0.0029)	(0.0032)
	0.950	0.6261	0.2966	0.2792	0.3538	0.1335	0.1622	0.1603	0.0746	0.1208
		(0.6313)	(0.0246)	(0.0369)	(1.1684)	(0.0088)	(0.0097)	(0.1135)	(0.0039)	(0.0045)
	0.975	0.8051	0.3885	0.3516	0.4879	0.1789	0.2109	0.2665	0.1016	0.1568
		(0.4516)	(0.0312)	(0.0527)	(0.3736)	(0.0118)	(0.0167)	(0.2234)	(0.0052)	(0.0077)
(4)	0.900	0.9162	0.8041	0.2226	0.4854	0.4510	0.1077	0.2876	0.2763	0.0697
		(0.2432)	(0.1532)	(0.0588)	(0.1093)	(0.0597)	(0.0089)	(0.0498)	(0.0247)	(0.0042)
	0.950	1.4972	1.2869	0.2725	0.9127	0.8092	0.1296	0.5585	0.5280	0.0812
		(0.4494)	(0.2337)	(0.0713)	(0.4895)	(0.1187)	(0.0142)	(0.1595)	(0.0554)	(0.0069)
	0.975	2.3371	1.9727	0.3331	1.5522	1.3387	0.1629	1.2223	0.9421	0.0995
		(1.0034)	(0.2835)	(0.0979)	(0.7483)	(0.1999)	(0.0248)	(1.4707)	(0.1110)	(0.0117)
(5)	0.900	0.0343	0.0091	0.0368	0.0298	0.0038	0.0315	0.0244	0.0021	0.0296
		(0.0224)	(0.0007)	(0.0013)	(0.0261)	(0.0002)	(0.0004)	(0.0238)	(0.0001)	(0.0002)
	0.950	0.1225	0.0110	0.0409	0.0351	0.0044	0.0345	0.0285	0.0023	0.0322
		(1.1145)	(0.0008)	(0.0020)	(0.0398)	(0.0003)	(0.0007)	0.0254	(0.0004)	(0.0004)
	0.975	0.0776	0.0135	0.0474	0.0455	0.0052	0.0397	0.0360	0.0027	0.0366
		(0.3266)	(0.0011)	(0.0034)	(0.0658)	(0.0003)	(0.0012)	(0.0309)	(0.0001)	(0.0006)

Table 3: average MSE and its standard deviation in brackets by 500 simulation runs for the simulation setting 1.

scenario	τ	n=20, p=100			n=50, p=150			n=100, p=200		
		BUP	TD	PEC	BUP	TD	PEC	BUP	TD	PEC
(1)	0.900	0.4484	0.2436	0.1988	0.2053	0.1077	0.1002	0.1109	0.0595	0.0660
		(0.2671)	(0.0195)	(0.0238)	(0.1273)	(0.0066)	(0.0058)	(0.0924)	(0.0030)	(0.0027)
	0.950	0.7021	0.314	0.2418	0.3681	0.1411	0.119	0.2075	0.0788	0.0761
		(0.4611)	(0.0246)	(0.0386)	(0.3066)	(0.0090)	(0.0091)	(0.2346)	(0.0039)	(0.0039)
	0.975	0.9218	0.4116	0.2945	0.5957	0.1890	0.1483	0.3364	0.1074	0.0925
		(0.5578)	(0.0312)	(0.0546)	(0.4751)	(0.0121)	(0.0152)	(0.3565)	(0.0053)	(0.0067)
(2)	0.900	0.7424	0.5427	0.3186	0.3560	0.2716	0.1502	0.2047	0.1549	0.0935
		(0.2933)	(0.1099)	(0.0762)	(0.1695)	(0.0728)	(0.0123)	(0.1886)	(0.0218)	(0.0050)
	0.950	1.1483	0.7855	0.3920	0.6656	0.4437	0.1832	0.3805	0.2684	0.1103
		(0.5078)	(0.1643)	(0.1096)	(0.6719)	(0.1658)	(0.0185)	(0.3563)	(0.0581)	(0.0075)
	0.975	1.7083	1.1095	0.4805	1.1714	0.7048	0.2342	0.6974	0.4648	0.1368
		(0.8614)	(0.1744)	(0.1493)	(0.9716)	(0.2652)	(0.0323)	(0.5981)	(0.1192)	(0.0126)
(3)	0.900	0.6616	0.4613	0.4093	0.2993	0.2041	0.2200	0.1684	0.1126	0.1540
		(0.2625)	(0.0392)	(0.0486)	(0.1163)	(0.0131)	(0.0134)	(0.1880)	(0.0058)	(0.0066)
	0.950	1.0027	0.5948	0.5229	0.4979	0.2675	0.2875	0.3031	0.1494	0.2042
		(0.5055)	(0.0495)	(0.0802)	(0.3671)	(0.0177)	(0.0215)	(0.4360)	(0.0077)	(0.0090)
	0.975	1.465	0.7811	0.6719	0.8605	0.3587	0.3831	0.5173	0.2036	0.2724
		(0.8018)	(0.0627)	(0.1154)	(0.8004)	(0.0237)	(0.0338)	(0.6708)	(0.0103)	(0.0156)
(4)	0.900	5.4073	5.2042	1.0318	3.3226	3.2871	0.4075	2.0358	2.0686	0.2295
		(2.1503)	(1.9812)	(0.9534)	(1.1548)	(1.0106)	(0.1258)	(0.6044)	(0.5259)	(0.1632)
	0.950	8.7171	8.0696	1.4256	6.5227	6.2094	0.5143	4.5541	4.4481	0.2939
		(2.8223)	(2.3418)	(1.4550)	(1.9576)	(1.5846)	(0.1540)	(1.4193)	(1.0287)	(0.3150)
	0.975	13.419	11.635	2.0054	11.202	9.8804	0.7372	8.9280	8.3663	0.3889
		(5.1223)	(1.6721)	(2.2733)	(4.0968)	(1.8550)	(0.5037)	(2.4679)	(2.7240)	(0.3161)
(5)	0.900	0.1135	0.0365	0.0572	0.0923	0.0153	0.0394	0.0561	0.0083	0.0333
		(0.0755)	(0.0027)	(0.0041)	(0.0878)	(0.0009)	(0.0011)	(0.0628)	(0.0004)	(0.0005)
	0.950	0.1430	0.0440	0.0651	0.1197	0.0177	0.0434	0.0896	0.0093	0.0356
		(0.1214)	(0.0034)	(0.0060)	(0.1033)	(0.0010)	(0.0018)	(0.0938)	(0.0005)	(0.0008)
	0.975	0.2489	0.0540	0.0769	0.1538	0.0209	0.0499	0.1145	0.0107	0.0396
		(0.6091)	(0.0042)	(0.0099)	(0.1272)	(0.0013)	(0.0031)	(0.1042)	(0.0006)	(0.0013)

Table 4: average MSE and its standard deviation in brackets by 500 simulation runs for the simulation setting 2.

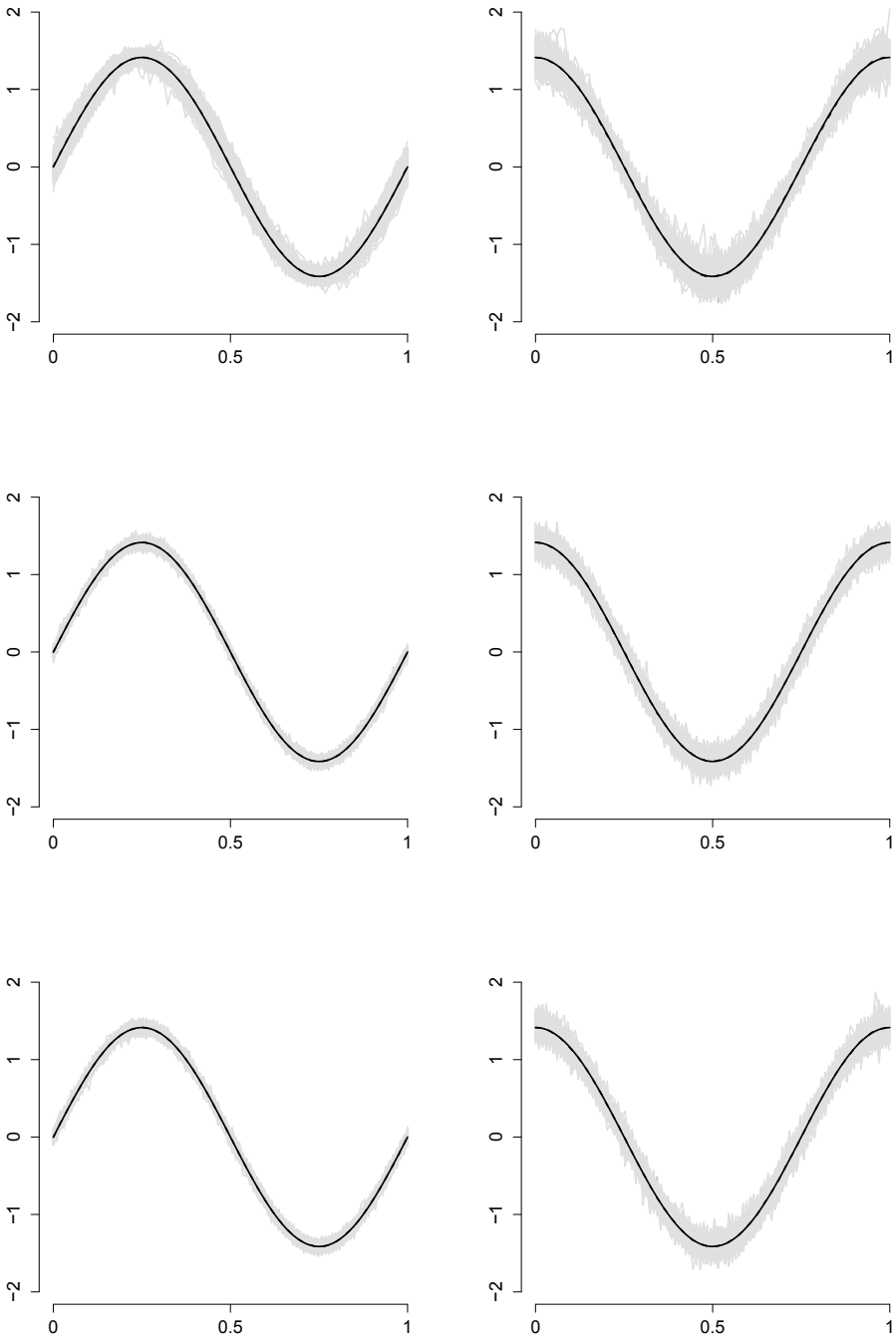


Figure 8: Estimated component functions (solid gray) by 500 simulation runs for simulation setting 1 scenario 1 small sample size and 95% expectile. The rows from the top to the bottom show respectively results produced by BUP, TD and PEC. Left panel corresponds to the first component function, right panel - to the second. The true functions are shown as solid black curves. The overall mean across simulation runs is shown as dashed black curve. The later can not be distinguished from the true curve.

PEC_algorithm_princdir
PEC_algorithm_topdown

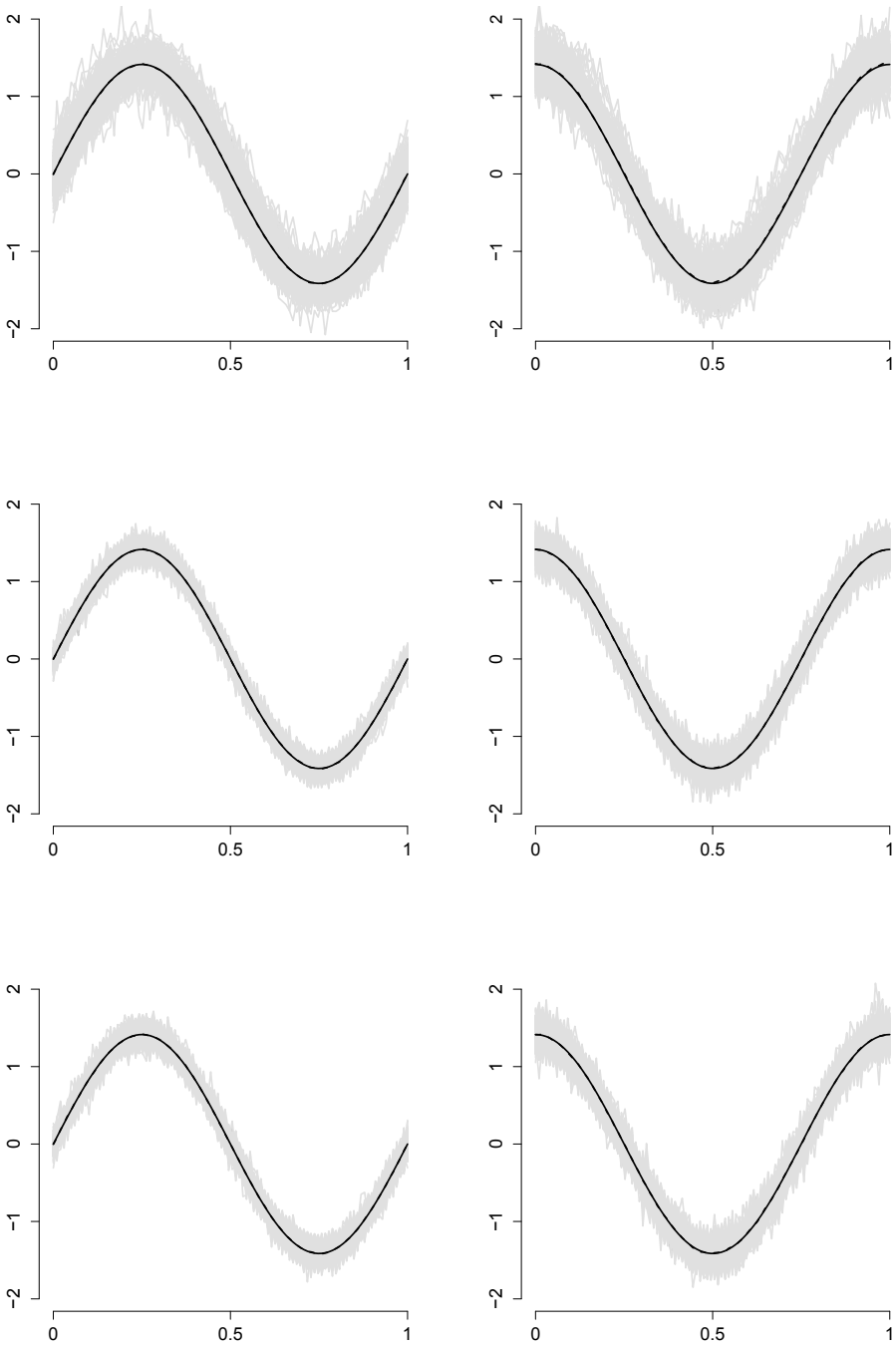


Figure 9: Estimated component functions (gray) by 500 simulation runs for simulation setting 2 scenario 1 small sample size and 95% expectile. The rows from the top to the bottom show respectively results produced by BUP, TD and PEC. Left panel corresponds to the first component function, right panel - to the second. The true functions are shown as solid black curves. The overall mean across simulation runs is shown as dashed black curve. The later can not be distinguished from the true curve.

PEC_algorithm_princdir
PEC_algorithm_topdown

References

Ashby, F. G. (2011): *Statistical Analysis of fMRI Data*, The MIT Press.

Burdejova, P., W. K. Härdle, P. Kokoszka, and Q. Xiong (in press 2016): "Change point and trend analysis of annual expestile curves of tropical storms," *Econometrics and Statistics*.

Cobzaş, Ş. (2013): *Functional analysis in asymmetric normed spaces*, Springer.

Daouia, A., S. Girard, and G. Stupfler (2016): "Estimation of Tail Risk based on Extreme Expectiles," Working paper or preprint.

Ferraty, F. and P. Vieu (2006): *Nonparametric Functional Data Analysis: Theory and Practice*, Springer.

Fraiman, R. and B. Pateiro-López (2012): "Quantiles for finite and infinite dimensional data," *Journal of Multivariate Analysis*, 108, 1–14.

Gao, Y., J. Xu, S. Yang, X. Tang, Q. Zhou, J. Ge, T. Xu, and R. Levinson (2014): "Cool roofs in China: Policy review, building simulations, and proof-of-concept experiments," *Energy Policy*, 74, 190 – 214.

Guo, M., L. Zhou, W. Härdle, and J. Huang (2015): "Functional data analysis of generalized regression quantiles," *Statistics and Computing*, 25, 189–202.

Horváth, L. and P. Kokoszka (2012): *Inference for Functional Data with Applications*, Springer.

Jolliffe, I. (2004): *Principal component analysis*, Springer.

Jorion, P. (2000): "Risk Management Lessons from Long-Term Capital Management," *European Financial Management*, 6, 277–300.

Klein Tank, A. M., F. W. Zwiers, and X. Zhang (2009): *Guidelines on Analysis of extremes in a changing climate in support of informed decisions for adaptation*, World Meteorological Organization.

Kong, L. and I. Mizera (2012): "Quantile tomography: using quantiles with multivariate data," *Statistica Sinica*, 22, 1589–1610.

Kuan, C.-M., J.-H. Yeh, and Y.-C. Hsu (2009): "Assessing value at risk with CARE, the Conditional Autoregressive Expectile models," *Journal of Econometrics*, 150, 261 – 270.

López Cabrera, B. and F. Schulz (2016): "Forecasting Generalized Quantiles of Electricity Demand: A Functional Data Approach," *Journal of the American Statistical Association*.

Majer, P., P. N. C. Mohr, H. R. Heekeren, and W. K. Härdle (2015): "Portfolio Decisions and Brain Reactions via the CEAD method," *Psychometrika*, 1–23.

Mohr, P. N., G. Biele, L. K. Krugel, S.-C. Li, and H. R. Heekeren (2010): "Neural foundations of risk-return trade-off in investment decisions," *NeuroImage*, 49, 2556 – 2563.

Mohr, P. N. and I. E. Nagel (2010): "Neural foundations of risk-return trade-off in investment decisions," *JNeurosci*, 30, 7755 – 7757.

Newey, W. and J. Powell (1987): "Asymmetric least squares estimation and testing," *Econometrica*, 819–847.

Ramsay, J. and B. Silverman (2005): *Functional data analysis*, Springer, New York.

Schnabel, S. (2011): "Expectile smoothing: new perspectives on asymmetric least squares. An application to life expectancy," Ph.D. thesis, Utrecht University.

Srebro, N. and T. Jaakkola (2003): "Weighted low-rank approximations," in *International Conference on Machine Learning*, 720–727.

Zhou, B., Y. Xu, J. Wu, S. Dong, and Y. Shi (2016): "Changes in temperature and precipitation extreme indices over China: analysis of a high-resolution grid dataset," *International Journal of Climatology*, 36, 1051–1066.

SFB 649 Discussion Paper Series 2016

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Downside risk and stock returns: An empirical analysis of the long-run and short-run dynamics from the G-7 Countries" by Cathy Yi-Hsuan Chen, Thomas C. Chiang and Wolfgang Karl Härdle, January 2016.
002 "Uncertainty and Employment Dynamics in the Euro Area and the US" by Aleksei Netsunajev and Katharina Glass, January 2016.
003 "College Admissions with Entrance Exams: Centralized versus Decentralized" by Isa E. Hafalir, Rustamdjan Hakimov, Dorothea Kübler and Morimitsu Kurino, January 2016.
004 "Leveraged ETF options implied volatility paradox: a statistical study" by Wolfgang Karl Härdle, Sergey Nasekin and Zhiwu Hong, February 2016.
005 "The German Labor Market Miracle, 2003 -2015: An Assessment" by Michael C. Burda, February 2016.
006 "What Derives the Bond Portfolio Value-at-Risk: Information Roles of Macroeconomic and Financial Stress Factors" by Anthony H. Tu and Cathy Yi-Hsuan Chen, February 2016.
007 "Budget-neutral fiscal rules targeting inflation differentials" by Maren Brede, February 2016.
008 "Measuring the benefit from reducing income inequality in terms of GDP" by Simon Voigts, February 2016.
009 "Solving DSGE Portfolio Choice Models with Asymmetric Countries" by Grzegorz R. Dlugoszek, February 2016.
010 "No Role for the Hartz Reforms? Demand and Supply Factors in the German Labor Market, 1993-2014" by Michael C. Burda and Stefanie Seele, February 2016.
011 "Cognitive Load Increases Risk Aversion" by Holger Gerhardt, Guido P. Biele, Hauke R. Heekeren, and Harald Uhlig, March 2016.
012 "Neighborhood Effects in Wind Farm Performance: An Econometric Approach" by Matthias Ritter, Simone Pieralli and Martin Odening, March 2016.
013 "The importance of time-varying parameters in new Keynesian models with zero lower bound" by Julien Albertini and Hong Lan, March 2016.
014 "Aggregate Employment, Job Polarization and Inequalities: A Transatlantic Perspective" by Julien Albertini and Jean Olivier Hairault, March 2016.
015 "The Anchoring of Inflation Expectations in the Short and in the Long Run" by Dieter Nautz, Aleksei Netsunajev and Till Strohsal, March 2016.
016 "Irrational Exuberance and Herding in Financial Markets" by Christopher Boortz, March 2016.
017 "Calculating Joint Confidence Bands for Impulse Response Functions using Highest Density Regions" by Helmut Lütkepohl, Anna Staszewska-Bystrova and Peter Winker, March 2016.
018 "Factorisable Sparse Tail Event Curves with Expectiles" by Wolfgang K. Härdle, Chen Huang and Shih-Kang Chao, March 2016.
019 "International dynamics of inflation expectations" by Aleksei Netšunajev and Lars Winkelmann, May 2016.
020 "Academic Ranking Scales in Economics: Prediction and Imdputation" by Alona Zharova, Andrija Mihoci and Wolfgang Karl Härdle, May 2016.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

SFB 649 Discussion Paper Series 2016

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

021 "CRIX or evaluating blockchain based currencies" by Simon Trimborn and Wolfgang Karl Härdle, May 2016.
022 "Towards a national indicator for urban green space provision and environmental inequalities in Germany: Method and findings" by Henry Wüstemann, Dennis Kalisch, June 2016.
023 "A Mortality Model for Multi-populations: A Semi-Parametric Approach" by Lei Fang, Wolfgang K. Härdle and Juhyun Park, June 2016.
024 "Simultaneous Inference for the Partially Linear Model with a Multivariate Unknown Function when the Covariates are Measured with Errors" by Kun Ho Kim, Shih-Kang Chao and Wolfgang K. Härdle, August 2016.
025 "Forecasting Limit Order Book Liquidity Supply-Demand Curves with Functional AutoRegressive Dynamics" by Ying Chen, Wee Song Chua and Wolfgang K. Härdle, August 2016.
026 "VAT multipliers and pass-through dynamics" by Simon Voigts, August 2016.
027 "Can a Bonus Overcome Moral Hazard? An Experiment on Voluntary Payments, Competition, and Reputation in Markets for Expert Services" by Vera Angelova and Tobias Regner, August 2016.
028 "Relative Performance of Liability Rules: Experimental Evidence" by Vera Angelova, Giuseppe Attanasi, Yolande Hiriart, August 2016.
029 "What renders financial advisors less treacherous? On commissions and reciprocity" by Vera Angelova, August 2016.
030 "Do voluntary payments to advisors improve the quality of financial advice? An experimental sender-receiver game" by Vera Angelova and Tobias Regner, August 2016.
031 "A first econometric analysis of the CRIX family" by Shi Chen, Cathy Yi-Hsuan Chen, Wolfgang Karl Härdle, TM Lee and Bobby Ong, August 2016.
032 "Specification Testing in Nonparametric Instrumental Quantile Regression" by Christoph Breunig, August 2016.
033 "Functional Principal Component Analysis for Derivatives of Multivariate Curves" by Maria Grith, Wolfgang K. Härdle, Alois Kneip and Heiko Wagner, August 2016.
034 "Blooming Landscapes in the West? - German reunification and the price of land." by Raphael Schoettler and Nikolaus Wolf, September 2016.
035 "Time-Adaptive Probabilistic Forecasts of Electricity Spot Prices with Application to Risk Management." by Brenda López Cabrera , Franziska Schulz, September 2016.
036 "Protecting Unsophisticated Applicants in School Choice through Information Disclosure" by Christian Basteck and Marco Mantovani, September 2016.
037 "Cognitive Ability and Games of School Choice" by Christian Basteck and Marco Mantovani, Oktober 2016.
038 "The Cross-Section of Crypto-Currencies as Financial Assets: An Overview" by Hermann Elendner, Simon Trimborn, Bobby Ong and Teik Ming Lee, Oktober 2016.
039 "Disinflation and the Phillips Curve: Israel 1986-2015" by Rafi Melnick and Till Strohsal, Oktober 2016.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

SFB 649 Discussion Paper Series 2016

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

040 "Principal Component Analysis in an Asymmetric Norm" by Ngoc M. Tran, Petra Burdejová, Maria Osipenko and Wolfgang K. Härdle, October 2016.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".